Latent Dirichlet Allocation (LDA)

Also Known As

Topic Modeling
The Domain: Natural Language Text

Collection of documents

Each document consists of a set of word tokens drawn (with replacement) from a set of word types

e.g., “The big dog ate the small dog.”

Goal

construct models of domain via unsupervised learning

i.e., learning structure of domain
What Does It Mean To Understand The Structure Of A Domain?

• Obtain a compact representation of each document

• Obtain a generative model that produces observed documents with high probability (and others with lower probability)
Two Contrasting Approaches To Modeling Environments Of Words And Text

Latent Semantic Analysis (LSA)

- mathematical model
- a bit hacky

Topic Model (LDA)

- probabilistic model
- principled -> has produced many extensions and embellishments
LSA

The set up

D documents
W distinct words
F = WxD cooccurrence matrix
$f_{wd}$ = frequency of word w in document d
LSA: Transforming The Co-occurrence Matrix

Relative entropy of a word across documents

\[ H_w = - \sum_{d=1}^{D} \frac{f_{wd}}{f_w} \log \left( \frac{f_{wd}}{f_w} \right) \]

\( f_{wd}/f_w : P(d|w) \)

\( H_w = \) value in [0, 1]
0=word appears in only 1 doc
1=word spread across all documents

Specificity: \((1-H_w)\)

0 = word tells you nothing about the document;
1 = word tells you a lot about the document
LSA: Transforming The Co-occurrence Matrix

\[ G = W \times D \text{ normalized coocurrence matrix} \]

\[ g_{wd} = \log\{f_{wd} + 1\}(1 - H_w) \]

- log transform common for word freq analysis
- +1 ensures no log(0)
- weighted by specificity

Representation of word i: row i of G

- problem: high dimensional representation
- problem: doesn’t capture similarity structure of documents
LSA: Representing A Word

Dimensionality reduction via SVD

\[ G = M_1 \quad M_2 \quad M_3 \]

\[ [WxD] = [WxR] [RxR] [RxD] \]

if \( R = \min(W,D) \) reconstruction is perfect

if \( R < \min(W,D) \) least squares reconstruction, i.e., capture whatever structure there is in matrix with a reduced number of parameters

Reduced representation of word \( i \): row \( i \) of \( (M_1 M_2) \)

Reduced representation of document \( j \): column \( j \) of \( (M_2 M_3) \)

Can used reduced representation to determine semantic relationships

What’s the advantage of a reduced representation?
LSA Versus Topic Model

The reduced representations in LSA are vectors whose elements (features)

• can be negative
• are completely unconstrained

If we wish to operate in a currency of probability, then the elements

• must be nonnegative
• must sum to 1

Terminology

• LSA = LSI = latent semantic indexing
• pLSI = probabilistic latent semantic indexing
• LDA

topic model
pLSI (Hoffman, 1999)

Probabilistic model of language production

Generative model

- Select a document with probability $P(D)$
- Select a (latent) topic with probability $P(Z|D)$
- Generate a word with probability $P(W|Z)$
- Produce pair $<d_i, w_i>$ on draw $i$

$$P(D, W, Z) = P(D) \ P(Z|D) \ P(W|Z)$$

$$P(D, W) = \sum_z P(D) \ P(z|D) \ P(W|z)$$

$$P(W | D) = \sum_z P(z|D) \ P(W|z)$$
Inferring Latent Variable

\[ P(Z|D,W) \]
\[ P(D, W, Z) = P(D) \ P(Z|D) \ P(W|Z) \]
\[ P(D, W) = \sum_z P(D) \ P(z|D) \ P(W|z) \]
\[ P(Z|D,W) = P(D, W, Z) / P(D, W) \]
\[ = P(Z|D) \ P(W|Z) / [\sum_z P(z|D) \ P(W|z)] \]
Plate Notation

Way of representing

- multiple documents
- multiple words per document
Plate Notation

Way of representing

- multiple documents
- multiple words per document
# Translating Notation

<table>
<thead>
<tr>
<th></th>
<th>Barber</th>
<th>Typical Topic Modeling Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>total # documents</td>
<td>$N$</td>
<td>$N$</td>
</tr>
<tr>
<td>total # topics</td>
<td>$K$</td>
<td>$T$</td>
</tr>
<tr>
<td>total # word types</td>
<td>$D$ (dictionary)</td>
<td>$W$</td>
</tr>
<tr>
<td>index over documents</td>
<td>$n$</td>
<td></td>
</tr>
<tr>
<td>index over words in document</td>
<td>$w$</td>
<td>$i$: index over document-word pairs</td>
</tr>
<tr>
<td>index over words in dictionary</td>
<td>$i$</td>
<td>${w_i, d_i}$</td>
</tr>
<tr>
<td>topic assignment</td>
<td>$\mathbf{Z}^n_w$</td>
<td>$z_i$: topic of word-document pair $i$</td>
</tr>
<tr>
<td>distribution over topics</td>
<td>${\pi^n_k}$</td>
<td>${\theta^d_{ji}}$</td>
</tr>
<tr>
<td>distribution over words</td>
<td>${\theta^k_i}$</td>
<td>${\phi^j_{wi}}$</td>
</tr>
<tr>
<td>index over topics</td>
<td>$k$</td>
<td>$j$</td>
</tr>
</tbody>
</table>
Two Approaches To Learning Conditional Probabilities

\[ P(Z=j \mid D=d_i) \text{ or } \theta_{j,d_i} \]
\[ P(W=w_i \mid Z=j) \text{ or } \phi_{w_i,j} \]

**Hoffmann (1999)**

Search for the single best \( \theta \) and \( \phi \) via gradient descent in cross entropy (difference between distribution) of data and model

\[- \sum_{w,d} n(d,w) \log P(d,w)\]

**Griffiths & Steyvers (2002, 2005); Blei, Ng, & Jordan (2003)**

Treat \( \theta \) and \( \phi \) as random variables.
Treating \( \theta \) And \( \phi \) As Random Variables

Can marginalize over uncertainty, i.e.,

\[
P(Z|D) = \int_{\theta} P(Z|D, \theta)P(\theta) \]

\[
P(W|Z) = \int_{\phi} P(W|Z, \phi)P(\phi) \]

Model
Treating $\theta$ And $\phi$ As Random Variables

The two conditional distributions are defined over *discrete alternatives*.

\[
P(Z=j \mid D=d_i) \text{ or } \theta_j^{d_i}
\]
\[
P(W=w_i \mid Z=j) \text{ or } \phi_{w_i}^j
\]

If $n$ alternatives, distribution can be represented by multinomial with $n-1$ degrees of freedom.

To represent $\theta$ and $\phi$ as random variables, need to encode a distribution over distributions...
Dirichlet Distribution

• represents probability distribution over multinomial distributions

You can think of the uncertainty space over \( n \) probabilities constrained such that \( P(x) = 0 \) if \( \sum x_i \neq 1 \) or if \( x_i < 0 \)...

...or the representational space over \( n-1 \) probabilities constrained such that \( P(x)=0 \) if \( \sum x_i > 1 \) or if \( x_i < 0 \).

• generalization of beta distribution

• for multinomial RV with \( n \) alternatives, Dirichlet has \( n \) parameters.

Each parameter is a count of the number of occurrences.

Why \( n \) and not \( n-1 \) since there are \( n-1 \) degrees of freedom?
Dirichlet Distribution (n=3)

1.00 1.00 1.00

1.00 1.00 10.00

1.00 5.00 1.00

2.00 1.00 1.00

2.00 3.00 5.00

20.00 30.00 50.00

5.00 4.00 1.00

0.90 1.00 0.80

0.50 0.50 0.50

p1 0.5 1

p2 0 0.5 1
Dirichlet Distribution (n=3)
Dirichlet Is Conjugate Prior Of Multinomial

Simple example

φ ~ Dirichlet(1, 3, 4)

O = {w1, w1, w2, w3, w2, w1}

φ | O ~ Dirichlet(4, 5, 5)

Weak assumption about prior

φ ~ Dirichlet(β, β, β)
Full Model

\[ \alpha \rightarrow \theta \rightarrow D \rightarrow Z \rightarrow W \rightarrow \phi \rightarrow \beta \]
Collapsed Gibbs Sampling Approach

1. Define Dirichlet priors on \( \theta^d_i \) and \( \phi^j \)

2. Perform sampling over latent variables \( Z \), integrating out or collapsing over \( \theta \) and \( \phi \)

\[
P(Z_i|Z_{-i}, D, W) \sim \int P(W|Z, \phi)P(Z|D, \theta)P(\phi)P(\theta)\,d\phi\,d\theta
\]

This can be done analytically due to Dirichlet-Multinomial relationship

Note: no explicit representation of posterior \( P(\theta, \phi| Z, D, W) \)
Collapsed Gibbs Sampling

\[ P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + \beta \cdot W} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(\cdot)} + T \alpha} \]

i: index over word-doc pairs

**Ignore \( \alpha \) and \( \beta \) for the moment**

- First term: proportion of topic j draws in which \( w_i \) picked
- Second term: proportion of words in document \( d_i \) assigned to topic j

This formula integrates out the Dirichlet uncertainty over the multinomial probabilities!

**What are \( \alpha \) and \( \beta \)?**

Effectively, they function as smoothing parameters

Large values -> more smoothing
Detailed Procedure For Sampling From $P(Z|D,W)$

1. Randomly assign each $<d_i, w_i>$ pair a $z_i$ value.

2. For each $i$, resample according to equation on previous slide (one iteration)

3. Repeat for a burn in of, say, 1000 iterations

4. Use current assignment as a sample and estimate $P(Z|D)$ and $P(W|Z)$

Typically with Gibbs sampling, the results of multiple chains (restarts) are used. Why wouldn’t that work here?
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services, Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Centers share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
<table>
<thead>
<tr>
<th>FEEL</th>
<th>MUSIC</th>
<th>BALL</th>
<th>SCIENCE</th>
<th>WORKERS</th>
<th>FORCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEELINGS</td>
<td>PLAY</td>
<td>GAME</td>
<td>STUDY</td>
<td>WORK</td>
<td>FORCES</td>
</tr>
<tr>
<td>FEELING</td>
<td>DANCE</td>
<td>TEAM</td>
<td>SCIENTISTS</td>
<td>LABOR</td>
<td>BODY</td>
</tr>
<tr>
<td>ANGRY</td>
<td>PLAYS</td>
<td>PLAY</td>
<td>SCIENTIFIC</td>
<td>JOBS</td>
<td>GRAVITY</td>
</tr>
<tr>
<td>WAY</td>
<td>STAGE</td>
<td>BASEBALL</td>
<td>KNOWLEDGE</td>
<td>WORKING</td>
<td>MASS</td>
</tr>
<tr>
<td>THINK</td>
<td>PLAYED</td>
<td>FOOTBALL</td>
<td>CHEMISTRY</td>
<td>WORKER</td>
<td>PULL</td>
</tr>
<tr>
<td>SHOW</td>
<td>BAND</td>
<td>PLAYERS</td>
<td>RESEARCH</td>
<td>WAGES</td>
<td>NEWTON</td>
</tr>
<tr>
<td>FEELS</td>
<td>AUDIENCE</td>
<td>GAMES</td>
<td>BIOLOGY</td>
<td>FACTORY</td>
<td>OBJECT</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>MUSICAL</td>
<td>PLAYING</td>
<td>MATHEMATICS</td>
<td>JOB</td>
<td>LAW</td>
</tr>
<tr>
<td>FRIENDS</td>
<td>DANCING</td>
<td>FIELD</td>
<td>LABORATORY</td>
<td>WAGE</td>
<td>DIRECTION</td>
</tr>
<tr>
<td>THINGS</td>
<td>RHYTHM</td>
<td>PLAYED</td>
<td>STUDYING</td>
<td>SKILLED</td>
<td>MOVING</td>
</tr>
<tr>
<td>MIGHT</td>
<td>PLAYING</td>
<td>PLAYER</td>
<td>SCIENTIST</td>
<td>PAID</td>
<td>REST</td>
</tr>
<tr>
<td>HELP</td>
<td>THEATER</td>
<td>COACH</td>
<td>PHYSICS</td>
<td>CONDITIONS</td>
<td>FALL</td>
</tr>
<tr>
<td>HAPPY</td>
<td>DRUM</td>
<td>BASKETBALL</td>
<td>FIELD</td>
<td>PAY</td>
<td>ACTING</td>
</tr>
<tr>
<td>FELT</td>
<td>ACTORS</td>
<td>SPORTS</td>
<td>STUDIES</td>
<td>FORCE</td>
<td>MOMENTUM</td>
</tr>
<tr>
<td>LOVE</td>
<td>SHOW</td>
<td>HIT</td>
<td>UNDERSTAND</td>
<td>MANY</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>ANGER</td>
<td>BALLET</td>
<td>BAT</td>
<td>STUDIED</td>
<td>HOURS</td>
<td>GRAVITATIONAL</td>
</tr>
<tr>
<td>BEING</td>
<td>ACTOR</td>
<td>TENNIS</td>
<td>SCIENCES</td>
<td>EMPLOYMENT</td>
<td>PUSH</td>
</tr>
<tr>
<td>WAYS</td>
<td>DRAMA</td>
<td>TEAMS</td>
<td>MANY</td>
<td>EMPLOYED</td>
<td>VELOCITY</td>
</tr>
<tr>
<td>FEAR</td>
<td>SONG</td>
<td>SOCCER</td>
<td></td>
<td>EMPLOYERS</td>
<td></td>
</tr>
</tbody>
</table>
Results

Wikipedia Topics
Relative Presence of Topics in all Documents

Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked for beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick’s films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that echoes structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionistic nature. A recurring theme in his films is man’s inhumanity to man. While often viewed as

{film, series, show}

words related documents related topics

film The X-Files
series Orson Welles
show Stanley Kubrick
character B movie
play Mystery Science Theater 3000
make Monty Python
episode Doctor Who
movie Sam Peckinpah
good Married... with Children
release History of film
feature The A-Team
television Pulp Fiction (film)
star Mad (magazine)

{theory, work, human}

words related documents related topics

theory Meme
work Intelligent design
human Immoral Kant
idea Philosophy of mathematics
term History of science
study Free will
view Truth

science Psychoanalysis
concept Charles Peirce
form Existentialism
world Deconstruction
argue Social sciences
social Idealism

(This and following slides from David Blei tutorial)
How might these graphs have been obtained?
How might this graph have been obtained?
Predicting word association norms

“the” -> ?

“dog” -> ?
Median Rank of k’th Associate

Note: multiple resamples can be used here
Combining Syntax and Semantics

LSA and Topic Model are “bag o’ words” models

Model sequential structure with 3d order HMM

- hidden state is category of word; 50 states
- 1 state for start or end of a sentence
- 48 states for document-independent words (syntax)
- 1 state for document-dependent words (semantics)

Semantics generated by topic model
### Multinomial distributions (most probable words in state)

<table>
<thead>
<tr>
<th>Syntax</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>MAP</td>
</tr>
<tr>
<td>YOU</td>
<td>DOCTOR</td>
</tr>
<tr>
<td>THEY</td>
<td>NORTH</td>
</tr>
<tr>
<td>I</td>
<td>EARTH</td>
</tr>
<tr>
<td>SHE</td>
<td>SOUTH</td>
</tr>
<tr>
<td>WE</td>
<td>POLE</td>
</tr>
<tr>
<td>IT</td>
<td>POLE</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>EQUATOR</td>
</tr>
<tr>
<td>EVERYONE</td>
<td>PATIENTS</td>
</tr>
<tr>
<td>OTHERS</td>
<td>WEST</td>
</tr>
<tr>
<td>SCIENTISTS</td>
<td>LINES</td>
</tr>
<tr>
<td>SOMEONE</td>
<td>DOCTORS</td>
</tr>
<tr>
<td>WHO</td>
<td>EAST</td>
</tr>
<tr>
<td>NOBODY</td>
<td>MEDICINE</td>
</tr>
<tr>
<td>ONE</td>
<td>WEST NURSE</td>
</tr>
<tr>
<td>SOMETHING</td>
<td>GLOBE</td>
</tr>
<tr>
<td>ANYONE</td>
<td>TREATMENT</td>
</tr>
<tr>
<td>EVERYBODY</td>
<td>POLES NURSES</td>
</tr>
<tr>
<td>SOME</td>
<td>PHYSICIAN</td>
</tr>
<tr>
<td>THEN</td>
<td>HOSPITALS</td>
</tr>
<tr>
<td></td>
<td>DR</td>
</tr>
<tr>
<td></td>
<td>LAND</td>
</tr>
<tr>
<td></td>
<td>SICK</td>
</tr>
<tr>
<td></td>
<td>ASSISTANT</td>
</tr>
<tr>
<td></td>
<td>COMPASS</td>
</tr>
<tr>
<td></td>
<td>EMERGENCY</td>
</tr>
<tr>
<td></td>
<td>CONTINENTS</td>
</tr>
<tr>
<td></td>
<td>PRACTICE</td>
</tr>
</tbody>
</table>