Distributional Semantics and Topic Modeling: Theory and Application

Baltic Summer School of Digital Humanities: Essentials of Coding and Encoding
Riga, July 2019

Christof Schöch
(Trier Center for Digital Humanities, Trier, Germany)
Overview

1. Introduction
2. Distributional Semantics: Principles and Methods
3. What are Word Embeddings?
4. What is Topic Modeling? Examples
5. Topic Models: the Theory
6. A Topic Modeling pipeline
7. First steps doing Topic Modeling
8. Advanced issues in Topic Modeling
9. Wrapping up
About this workshop

- Slides available online: https://christofs.github.io/riga/#/
- Download code and sample datasets: https://github.com/dh-trier/topicmodeling
About this workshop

- Context, examples, theory, demo, hands-on for Topic Modeling

- Slides available online: https://christofs.github.io/riga/#/
- Download code and sample datasets: https://github.com/dh-trier/topicmodeling
About this workshop

- Context, examples, theory, demo, hands-on for Topic Modeling
- Python-based, but not a Python workshop
  ("read and run" code, rather than write code)

---

- Slides available online: https://christofs.github.io/riga/#
- Download code and sample datasets:
  https://github.com/dh-trier/topicmodeling
About this workshop

- Context, examples, theory, demo, hands-on for Topic Modeling
- Python-based, but not a Python workshop ("read and run" code, rather than write code)
- Learning goal: you understand how a Topic Model is created and can run your own Topic Modeling Pipeline

- Slides available online: https://christofs.github.io/riga/##/
- Download code and sample datasets: https://github.com/dh-trier/topicmodeling
About myself

- Professor of Digital Humanities
- Not a computer scientist, not a statistician
- French literary scholar by training
- Interests in corpus building and quantitative text analysis
- see: https://christof-schoech.de/en
About you: raise your hand if...

• ... you are a literary scholar
About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist
About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist
- ... you are a digital humanist
About you: raise your hand if...

• ... you are a literary scholar
• ... you are a historian
• ... you are a sociologist
• ... you are a (computational / corpus) linguist
• ... you are a computer scientist
• ... you are a digital humanist
• ... you are a librarian
About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist
- ... you are a digital humanist
- ... you are a librarian
- ... you consider yourself to be a local
Distributional Semantics: Principles and Methods
Basic intuition about distributional semantics

- "Her friend's ...... was located on the second floor of the house."
Basic intuition about distributional semantics

- "Her friend's ...... was located on the second floor of the house."
- "apartment"!
Basic intuition about distributional semantics

- "Her friend's ...... was located on the second floor of the house."
- "apartment"!
- "room"!
Basic intuition about distributional semantics

- "Her friend's ...... was located on the second floor of the house."
- "apartment"!
- "room"!
- "balcony"?
Basic intuition about distributional semantics

- "Her friend's ...... was located on the second floor of the house."
- "apartment"!
- "room"!
- "balcony"?
- "cat"??
Basic intuition about distributional semantics

- "Her friend's ...... was located on the second floor of the house."
- "apartment"!
- "room"!
- "balcony"?
- "cat"??
- "shark"???
What does this example tell us?

- We are able to rank the likelihood of these words in the given context
- We use world knowledge, but also linguistic competency, for this
- Computers can learn this too, based on cooccurrence patterns
What does this example tell us?

- We are able to rank the likelihood of these words in the given context
- We use world knowledge, but also linguistic competency, for this
- Computers can learn this too, based on cooccurrence patterns
- That's how distributional semantics works!
Basic idea

- The meaning of words depends on their context
  "You shall know a word by the company it keeps" (Firth, 1957)
Basic idea

- The meaning of words depends on their context
  "You shall know a word by the company it keeps" (Firth, 1957)
- Words frequently appearing in similar contexts have similar meanings
- Words that can appear in very similar, specific contexts have similar grammatical functions
Two applications of this idea

- Topic Modeling
- Word Embeddings
What are Word Embeddings?
Information Retrieval: Vector Space Model
Information Retrieval: Vector Space Model

- Each document has a certain place in a vector space
- That place is determined by the keywords that appear in the document
- Each word is a dimension in the vector space
- Documents with shared vocabulary end up in the same area of the vector space
Information Retrieval: Vector Space Model

(Image Credit: Riclas, Wikipedia, Creative Commons Attribution 3.0)
Words in vector space

(Artificial data. Image credit: Christof Schöch, 2019. Creative Commons Attribution 4.0 International)
Example: French Wikipedia Model

- 1.8 million articles, 750 million words
- transform term-document-matrix into dense matrix
- "low-dimensional", dense representation
- skip-gram model, 300 dimensions
- vector semantics: geometric relations = semantic relations
Similar Words Query

Query: ['poésie_nom', 10]
Result: poétique_adj 0.841
poème_nom 0.790
prose_nom 0.733
littérature_nom 0.715
poète_nom 0.704
poétique_nom 0.701
poésie_nom 0.700
anthologie_nom 0.695
littéraire_adj 0.655
sonnet_nom 0.651

(authentic data, Wikipedia model)
Similarity Query

Query: ['prose_nom', 'littérature_nom']
Result: 0.511518681366

Query: ['poésie_nom', 'littérature_nom']
Result: 0.714615326722

(authentic data, Wikipedia model)
Evaluation

- Method: Using a "find-the-wrong word"-task
Evaluation

- Method: Using a "find-the-wrong word"-task
- Lists of similar words:
  - vert, bleu, jaune, rouge, orange
  - billet, monnaie, portemonnaie, payement
Evaluation

• Method: Using a "find-the-wrong word"-task
• Lists of similar words:
  ▪ vert, bleu, jaune, rouge, orange
  ▪ billet, monnaie, portemonnaie, payement
• Generate lists with an error
  ▪ vert, bleu, monnaie, jaune, rouge
Evaluation

- Method: Using a "find-the-wrong word"-task
- Lists of similar words:
  - vert, bleu, jaune, rouge, orange
  - billet, monnaie, portemonnaie, payement
- Generate lists with an error
  - vert, bleu, monnaie, jaune, rouge
- Wikipedia model: 90% accuracy in finding the error
Axes of meaning

(Artificial data. Image credit: Christof Schöch, 2019, Creative Commons Attribution 4.0 International)
Axes of meaning (Ryan Heuser)

Axis query

Axis: [['bonheur', 'joie'], 'malheur', 'tristesse']] # positive
# negative

Query: ange
Result: 0.0875

Query: monstre
Result: -0.1407

(authentic data)
Time for questions!
References

What is Topic Modeling?
(a) Some fundamentals
Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
- Each document is understood as a mixture of topics
Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
- Each document is understood as a mixture of topics
- The purpose is to discover thematic trends and patterns
Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
- Each document is understood as a mixture of topics
- The purpose is to discover thematic trends and patterns
- Discovered through generative probabilistic modeling
Usage scenarios
Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields
Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields
- Recommender Systems: Recommend similar journal articles etc. to users
Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields
- Recommender Systems: Recommend similar journal articles etc. to users
- Exploration of text collections: what is an email or newspaper corpus about?
Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields
- Recommender Systems: Recommend similar journal articles etc. to users
- Exploration of text collections: what is an email or newspaper corpus about?
- Research questions from literary studies, cultural studies, history of ideas: topics across authors, genres, time periods
Explorative Visualization

Signs at 40: http://signsat40.signsjournal.org/topic-model/#/model/grid

<table>
<thead>
<tr>
<th>Topic</th>
<th>%</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>India: India, Indian, class, state, hindu, caste, nation, middle</td>
<td>45.5</td>
<td>1836</td>
</tr>
<tr>
<td>Bodies: body, subject, bodies, space, time, history, object, relation</td>
<td>11.1</td>
<td>450</td>
</tr>
<tr>
<td>The social: social, power, experience, important, part, ways, terms</td>
<td>10.6</td>
<td>431</td>
</tr>
<tr>
<td>Film: film, films, cinema, narrative, work, desire, scene, cultural</td>
<td>7.5</td>
<td>305</td>
</tr>
<tr>
<td>Gender theory: gender, men, masculinity, masculine, gendered, sex</td>
<td>5.2</td>
<td>211</td>
</tr>
<tr>
<td>Media images: public, beauty, media, images, nation, popular, femininity, fashion</td>
<td>4.5</td>
<td>185</td>
</tr>
<tr>
<td>Lesbian, gay, queer: lesbian, gay, lesbians, queer, community,</td>
<td>3.4</td>
<td>140</td>
</tr>
<tr>
<td>Women’s roles: women, men, female, woman, male, society, role, status</td>
<td>3.3</td>
<td>135</td>
</tr>
<tr>
<td>China: Chinese, China, state, long, western, taiwan, cultural, li,</td>
<td>1.8</td>
<td>74</td>
</tr>
<tr>
<td>Globalization: world, global, states, international, united, economic</td>
<td>1.8</td>
<td>74</td>
</tr>
<tr>
<td>Asian, American, Caribbean: Asian, American, caribbean, cuba,</td>
<td>1.5</td>
<td>62</td>
</tr>
<tr>
<td>Sexuality: sexual, sexuality, sex, female, desire, love, male,</td>
<td>1.1</td>
<td>45</td>
</tr>
<tr>
<td>Family, poverty, welfare: family, children, families, marriage,</td>
<td>1.1</td>
<td>43</td>
</tr>
<tr>
<td>Performance: music, performance, girls, play, dance, girl,</td>
<td>0.7</td>
<td>27</td>
</tr>
<tr>
<td>Law: law, rights, legal, state, court, discrimination, laws, act,</td>
<td>0.4</td>
<td>17</td>
</tr>
<tr>
<td>Political movements: political, movement, politics, national, members</td>
<td>0.4</td>
<td>17</td>
</tr>
<tr>
<td>Rural economics: labor, production, rural, household, development</td>
<td>0.2</td>
<td>10</td>
</tr>
<tr>
<td>Feminist movements: feminist, feminism, women, feminists</td>
<td>0.1</td>
<td>6</td>
</tr>
<tr>
<td>Role</td>
<td>Movie</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>McCormick, Richard W. &quot;From 'Coliger' to Dietrich: Sexual, Social, and Cinematic Discourses in Weimar Film.&quot; Signs 18, no. 3 (Spring 1993): 649–668.</td>
<td>1903</td>
<td></td>
</tr>
<tr>
<td>Williams, Linda. &quot;Why I Did Not Want to Write This Essay.&quot; Signs 30, no. 1 (Autumn 2004): 1254–800.</td>
<td>433</td>
<td></td>
</tr>
</tbody>
</table>
Prominent topics for movie

Click row labels to go to the corresponding topic page; click a word to show the topic list for that word.

Film
Existing Studies
Existing Studies

Existing Studies

- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
Existing Studies

- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
Existing Studies

- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Matthew Jockers, Macroanalysis (2013): novel, nationality, gender
Existing Studies

- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Matthew Jockers, Macroanalysis (2013): novel, nationality, gender
- Ben Schmidt: "Typical TV episodes" (2014): TV shows; temporal development
Existing Studies

- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Matthew Jockers, Macroanalysis (2013): novel, nationality, gender
- Ben Schmidt: "Typical TV episodes" (2014): TV shows; temporal development
- Christof Schöch, "Topic Modeling Genre" (2017): drama, subgenres
(b) A topic model for French crime fiction
Text collection: 840 French Novels

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)
Crime fiction (prototypical)
Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
- Plot: violent crime, rational elucidation
Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
- Plot: violent crime, rational elucidation
- Setting: urban space
Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
- Plot: violent crime, rational elucidation
- Setting: urban space
- => Hypotheses regarding possible topics
Topic and subgenre

Topic 10: detective, inspector, police

Distinctive of crime fiction (content & statistics) ($p < \alpha = 0.01$)  
(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)
Topic and subgenre

Topic 49: death, crime, to kill

Distinctive of crime fiction (content & statistics) (p < α=0.01)

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int’l)
Topic 47: door, room, to open

Statistically distinctive ($p < \alpha = 0.01$); but content-wise?

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)
Topic and subgenre

Topic 26: beach, sand, sun

Distinctive of non-criminal fiction ($p < \alpha = 0.001$)

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)
Topics over text segments

Topic 2: judge, prison, lawyer/attorney

Statistically significant (crime fiction): (1,4), (4,5) etc.

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)
Topics over text segments

**Topic 33: black, hair, eyes, wear, eye, face**

Statistically significant: crime fiction all but (2,3); non-crime fiction (1,3), (2,5)

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)
Overall results
Overall results

• A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
Overall results

• A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
• Topics are not just themes, but also narrative motives, descriptive elements, character sets
Overall results

- A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
- Topics are not just themes, but also narrative motives, descriptive elements, character sets
- Textual progression: only a few topics have significant trends
Overall results

- A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
- Topics are not just themes, but also narrative motives, descriptive elements, character sets
- Textual progression: only a few topics have significant trends
- Overall: we can detect thematic trends in 840 novels without reading (all of) them!
Time for questions
Topic Modeling: Theory
(a) How does a topic model look like?
On a practical level

- A topic is a group of words with some (semantic) relation (e.g., common theme, motive, etc.)
On a practical level

- A topic is a group of words with some (semantic) relation (e.g., common theme, motive, etc.)
- Each topic is made up of words of varying importance and relevance to the topic
On a practical level

- A topic is a group of words with some (semantic) relation (e.g., common theme, motive, etc.)
- Each topic is made up of words of varying importance and relevance to the topic
- Each document is made up of several topics in various proportions
On a technical level

- A topic model is an abstract representation of all topics and documents in a collection
On a technical level

- A topic model is an abstract representation of all topics and documents in a collection
- A topic is a probability distribution over words
On a technical level

- A topic model is an abstract representation of all topics and documents in a collection
- A topic is a probability distribution over words
- A document is a probability distribution over topics
On a technical level

- A topic model is an abstract representation of all topics and documents in a collection
- A topic is a probability distribution over words
- A document is a probability distribution over topics
- The Dirichlet distribution (in LDA) describes the topic mixture distribution of the model
Dirichlet distributions

Describe the topic mixture distributions of the model. Here several possible distributions with three topics.

Draws from a 3-dimensional Dirichlet with different $\alpha$

- $\alpha = (1, 1, 1)$
- $\alpha = (2, 2, 2)$
- $\alpha = (10, 10, 10)$
- $\alpha = (0.5, 0.5, 0.5)$
- $\alpha = (0.1, 0.1, 0.1)$
- $\alpha = (10, 3, 5)$
Words in topic distribution

<table>
<thead>
<tr>
<th>topic</th>
<th>word</th>
<th>rank</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>horse</td>
<td>1</td>
<td>169.0182333</td>
</tr>
<tr>
<td>5</td>
<td>road</td>
<td>2</td>
<td>120.0182333</td>
</tr>
<tr>
<td>5</td>
<td>carriage</td>
<td>3</td>
<td>68.01823332</td>
</tr>
<tr>
<td>5</td>
<td>sergeant</td>
<td>4</td>
<td>52.01823332</td>
</tr>
<tr>
<td>5</td>
<td>companion</td>
<td>5</td>
<td>40.01823332</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>bridle</td>
<td>50</td>
<td>9.018233317</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>charge</td>
<td>100</td>
<td>4.018233317</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>garden</td>
<td>500</td>
<td>0.018233317</td>
</tr>
</tbody>
</table>

(Each word has a score in each topic; here ordered by topic/rank)
Topics in document distribution

<table>
<thead>
<tr>
<th>doc</th>
<th>doc-id</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>acd009§0005.txt</td>
<td>0.0009913736</td>
<td>0.0003037892</td>
<td>0.0002364825</td>
<td>0.0980390649</td>
</tr>
<tr>
<td>1</td>
<td>acd005§0068.txt</td>
<td>0.0015847905</td>
<td>0.0004856314</td>
<td>0.0007028928</td>
<td>0.1700251167</td>
</tr>
<tr>
<td>2</td>
<td>acd010§0026.txt</td>
<td>0.1416548293</td>
<td>0.2319451602</td>
<td>0.0001779831</td>
<td>0.0236859739</td>
</tr>
<tr>
<td>3</td>
<td>acd011§0020.txt</td>
<td>0.0007777962</td>
<td>0.0002383421</td>
<td>0.0001855357</td>
<td>0.0246910723</td>
</tr>
<tr>
<td>4</td>
<td>acd005§0048.txt</td>
<td>0.0813954255</td>
<td>0.0002265121</td>
<td>0.2049190768</td>
<td>0.1072239384</td>
</tr>
<tr>
<td>5</td>
<td>acd006§0001.txt</td>
<td>0.0007063237</td>
<td>0.0002164406</td>
<td>0.0001684866</td>
<td>0.1617409925</td>
</tr>
<tr>
<td>6</td>
<td>acd006§0070.txt</td>
<td>0.0261032958</td>
<td>0.0002216979</td>
<td>0.0001725791</td>
<td>0.0928001193</td>
</tr>
<tr>
<td>7</td>
<td>acd004§0054.txt</td>
<td>0.0047744552</td>
<td>0.0002815341</td>
<td>0.0002191581</td>
<td>0.3144879413</td>
</tr>
<tr>
<td>8</td>
<td>acd010§0036.txt</td>
<td>0.2589537644</td>
<td>0.0001861433</td>
<td>0.0116724568</td>
<td>...</td>
</tr>
<tr>
<td>9</td>
<td>acd005§0031.txt</td>
<td>0.048476437</td>
<td>0.0002322766</td>
<td>0.000180814</td>
<td>0.1767559653</td>
</tr>
<tr>
<td>10</td>
<td>acd006§0019.txt</td>
<td>0.0130915999</td>
<td>0.0002951922</td>
<td>0.0083153112</td>
<td>0.0993073862</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

(Each topic has a score in each document; ordered by document)
(b) How is a Topic Model created?
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
Some relevant ideas

• The most widespread implementation uses 'Latent Dirichlet Allocation'
• Follows the "bag-of-words"-model: word order is irrelevant
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
- Discovers words which frequently occur together or in similar contexts (=topics)
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
- Discovers words which frequently occur together or in similar contexts (=topics)
- Infers how important each word is in each topic
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
- Discovers words which frequently occur together or in similar contexts (=topics)
- Infers how important each word is in each topic
- Infers how important each topic is in each document
Generative, inverted, iterative

"A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents."

(Steyvers and Griffiths 2006)
Inference problem: observed data

Inferred, latent model

The starting point of LDA

• We have the documents with their words (e.g. as a word/document frequency matrix)
The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)
- We are looking for the word distributions per topic, the topic distributions per document, and the topic assignment of each word
The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)
- We are looking for the word distributions per topic, the topic distributions per document, and the topic assignment of each word
- Both distributions are dependent on each other (if a topic changes, the topic distributions change)
The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)
- We are looking for the word distributions per topic, the topic distributions per document, and the topic assignment of each word
- Both distributions are dependent on each other (if a topic changes, the topic distributions change)
- And both distributions need to fit with the original documents
The generative model behind LDA

- For each topic, there is a distribution over words
- For each document, there is a distribution over topics
- For each word in each document:
  - We sample a topic from the topic distribution of that document
  - We sample a word from the word distribution of that topic
- This can only work if we have the distributions; which we don't
Random initialization

- For each document, we generate a random distribution over topics
Random initialization

- For each document, we generate a random distribution over topics
- For each topic, we generate a random distribution over words
Random initialization

- For each document, we generate a random distribution over topics
- For each topic, we generate a random distribution over words
- For each word in each document:
  - Sample a topic from the topic distribution
  - Sample a word from the word distribution of that topic
Random initialization

- For each document, we generate a random distribution over topics
- For each topic, we generate a random distribution over words
- For each word in each document:
  - Sample a topic from the topic distribution
  - Sample a word from the word distribution of that topic
- Now we have a model; but we know it's most likely wrong (=low confidence)
Inference: iterative approximation

- Using the observed data and our (random/erroneous) model, we can improve the model
Inference: iterative approximation

- Using the observed data and our (random/erroneous) model, we can improve the model
- One among several methods: Gibbs sampling
Inference: iterative approximation

• Using the observed data and our (random/erroneous) model, we can improve the model
• One among several methods: Gibbs sampling
  ▪ For one word in one document, remove the existing topic assignment
  ▪ Based on the model and the other words in the document, assign a new topic to the word (cooccurrence!)
  ▪ Update the overall model according to this assignment;
• Repeat until your time runs out or your evaluation task says it's ok to stop
How it works exactly, clearly explained

"I think you should be more explicit in step two..."

For reasons of copyright restrictions, please see here: http://www.sciencecartoonsplus.com/gallery/math/index.php
Time for questions