Latent Dirichlet Allocation

Daniel Lawson — University of Bristol

Lecture 07.1.3 (v1.0.2)

Signposting

- This lecture combines knowledge from 7.1.1 on topic models with 7.1.2 on Bayesian Methodology to describe one of the most successful tools in natural language processing.
- In 7.2 we cover some practicalities.
- In the workshop we implement these models in practice.

ILOs

- ILO2 Be able to use and apply basic machine learning tools
- ILO3 Be able to make and report appropriate inferences from the results of applying basic tools to data

Beyond the bag of words

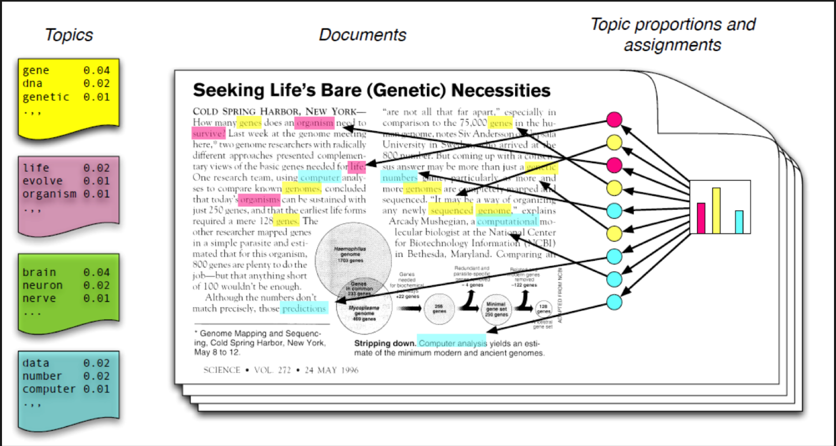
- The Bag-of-words is a **vector representation** of a set of documents.
  - i.e. a feature space embedding.
- But how can we use this? How do we **compare** documents?
  - We could perform **dimensionality reduction** via PCA,
  - **Distance metrics** such as Cosine Similarity,
  - etc.
- Or we can model the similarity. The most successful approach for this is **Latent Dirichlet Allocation** (LDA).

Modelling a Bag Of Words using Latent Dirichlet Allocation

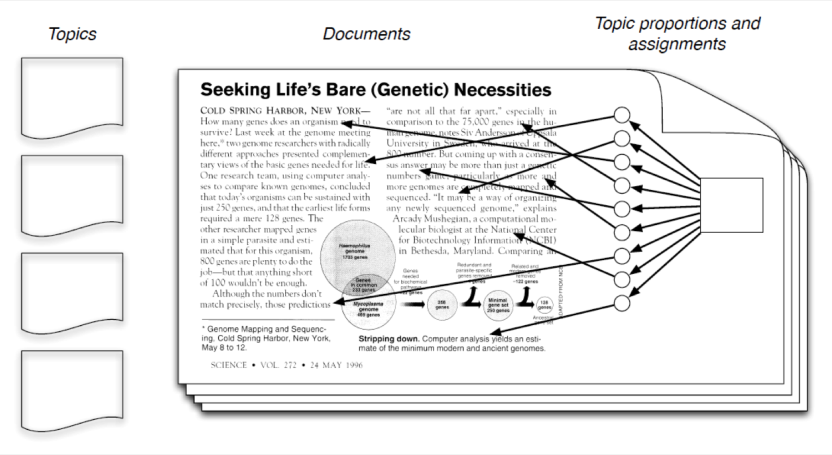
- Each document is modelled as a **mixture** of topics,
- Each topic is modelled as a **distribution** over words,
- Some Bayesian modelling magic allows the documents to be a theoretically **infinite mixture**,
- With content from [1] which also contains cyber examples (without data).

[1] Topic Modeling and Latent Dirichlet Allocation: An Overview (Weifeng Li, Sagar Samtani and Hsinchun Chen)

LDA Motivation - The setup

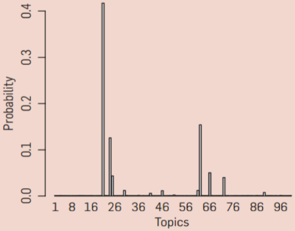


LDA Motivation - Data in Practice



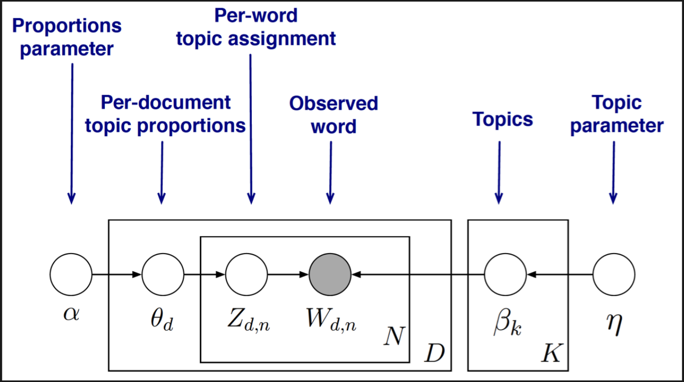
LDA Motivation - Example

The resulting output from an LDA model would be sets of topics containing keywords which would then be manually labeled. On the left are the inferred topic proportions for the example article from the pervious figure.



"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

LDA Probabilistic Graphical Model



This is **plate notation** for **Bayesian Graphical Models**.

LDA Definition

- The overall **word distribution** is η, an N-vector.
- The overall **topic distribution** is α, a K-vector.
- Each **topic** k is described by a **word frequency** vector $\beta_k \sim \mathrm{Dirichlet}(\eta)$.
- Each **document** d is described by a **topic frequency** vector $\theta_d \sim \mathrm{Dirichlet}(\alpha)$.
- When generating word i from document d, we **generate a topic** $z_{di} \sim \mathrm{Multinomial}(\theta_d)$.
- And then **generate a word** $w_{di} \sim \mathrm{Multinomial}(\beta_d)$.

LDA properties

- Because it is a generative model, we can can ask it to simulate documents.
- These approaches are embarrassing:
  - in the sense that if you simulate from the model, it generates garbage,
  - because words are independent.
- They should be thought of instead as keyword generators.
- This is extremely useful for a variety of text categorisation tasks.
- It can operate:
  - supervised (where we insist that some documents have pre-defined topic distributions) or
  - unsupervised (where nothing is assumed apriori about topics).

LDA implementation

- LDA implementations[2] use a conjugate model (Multinomial distribution is conjugate to the Dirichlet prior).
- It uses Variational Bayes to write the problem as an optimisation problem.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

Further notes on LDA

- LDA models a matrix $Y = AX$, where:
  - Y is the data (N rows containing L word frequencies),
  - X are the topics (K rows containing L word frequencies) and
  - A is a mixture, (N rows containing K topics)
- This is a common problem called **matrix decomposition**.
- What makes LDA special is that **words are sparse**, meaning that there are many words but most words don't appear in most documents.
- You can run LDA on any problem of this type, but there are other approaches for dense data. (We return to **sparsity** later.)

Extensions

- We will not cover them, but if you work with document models you may want a more realistic model.
- Predictive text uses Markov Chains to predict $p(t(i)|d, t(i-1))$.
- Neural Networks generate arbitrary correlation structure, e.g.
  - Mathgen generates random papers,
  - Topic-RNN infers a topic model using a Neural Network.

Reflection

- What is a bag of words, conceptually?
- What are the advantages of LDA over Bag of words?
- And vice-versa?
- Could you use SVD on a bag of words?
- Why would we use either, when empirical accuracy of neural-network approaches is higher?

Signposting

- Next: Practicalities of text modelling.
- References:
  - Topic Modeling and Latent Dirichlet Allocation: An Overview (Weifeng Li, Sagar Samtani and Hsinchun Chen)
  - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation", Journal of machine Learning research 3.Jan (2003): 993-1022.