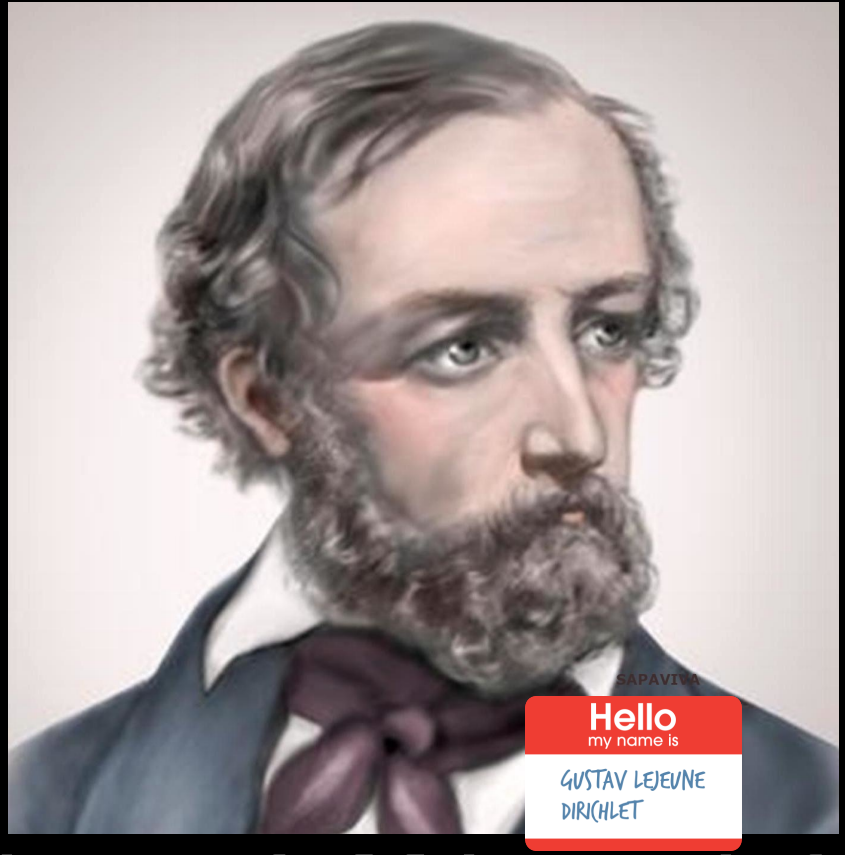


Topic Modelling with LDA

Instructor: Xandra Dave Cochran
April 3-10, 2024
Centre for Data, Culture & Society

Topic Modelling:

- Unsupervised Machine Learning

Topic Modelling:

- Unsupervised Machine Learning
- Identifies clusters of related words in text

Topic Modelling:

- Unsupervised Machine Learning
- Identifies clusters of related words in text
- Does not require predefined categories – good for discovery and exploration of a dataset

Introductions!

Is there anyone here today that wasn't here for BERTopic last week? If so – HELLO! Welcome! And…

Introductions!

Is there anyone here today that wasn't here for BERTopic last week? If so – HELLO! Welcome! And…

What is your previous experience with machine learning?

Why are you interested in topic modelling?

Have you used LLMs before?

Is there a dataset you have in mind to use for topic modelling in future?

Latent Dirichlet Allocation

- This is a rather more old-school model than BERTopic – introduced by Blei, Ng & Jordan in 2003

Latent Dirichlet Allocation

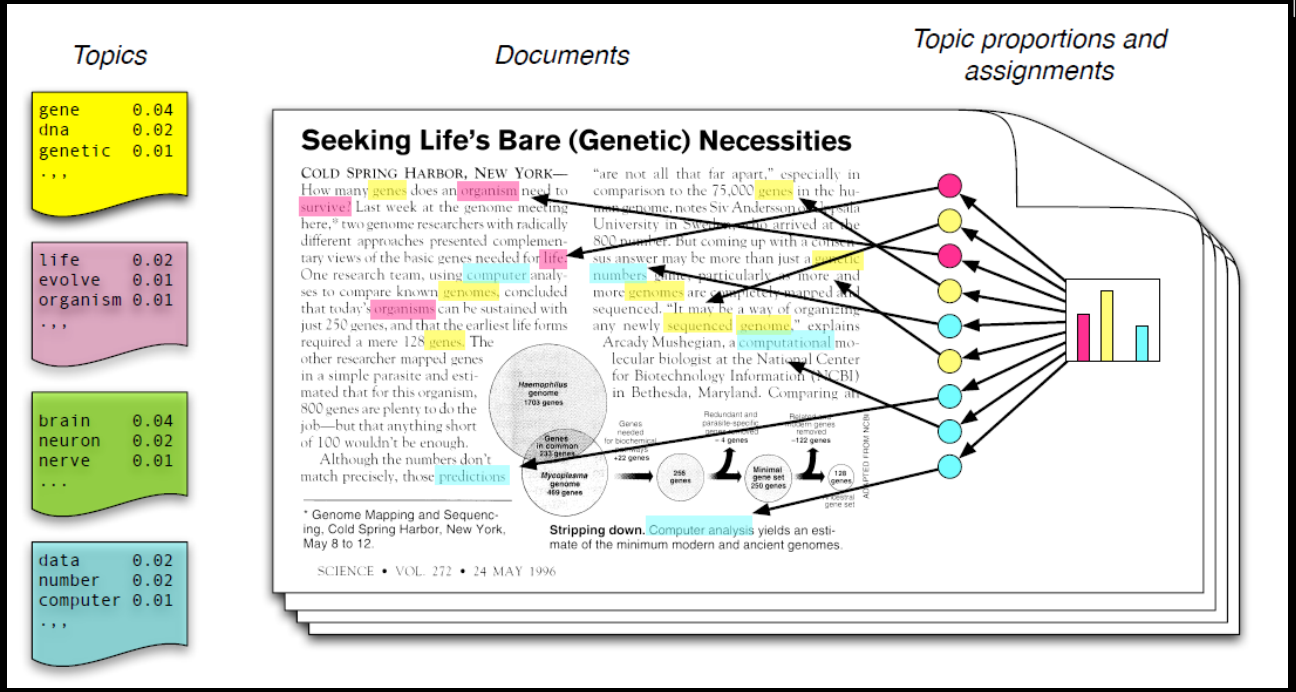
- This is a rather more old-school model than BERTopic – introduced by Blei, Ng & Jordan in 2003
- No billion-parameter neural networks here: the model is simpler & and, arguably, it’s easier to interpret its outputs (eXplainable AI – XAI)

Latent Dirichlet Allocation

- This is a rather more old-school model than BERTopic – introduced by Blei, Ng & Jordan in 2003
- No billion-parameter neural networks here: the model is simpler & and, arguably, it’s easier to interpret its outputs (eXplainable AI – XAI)
- But is it as good?

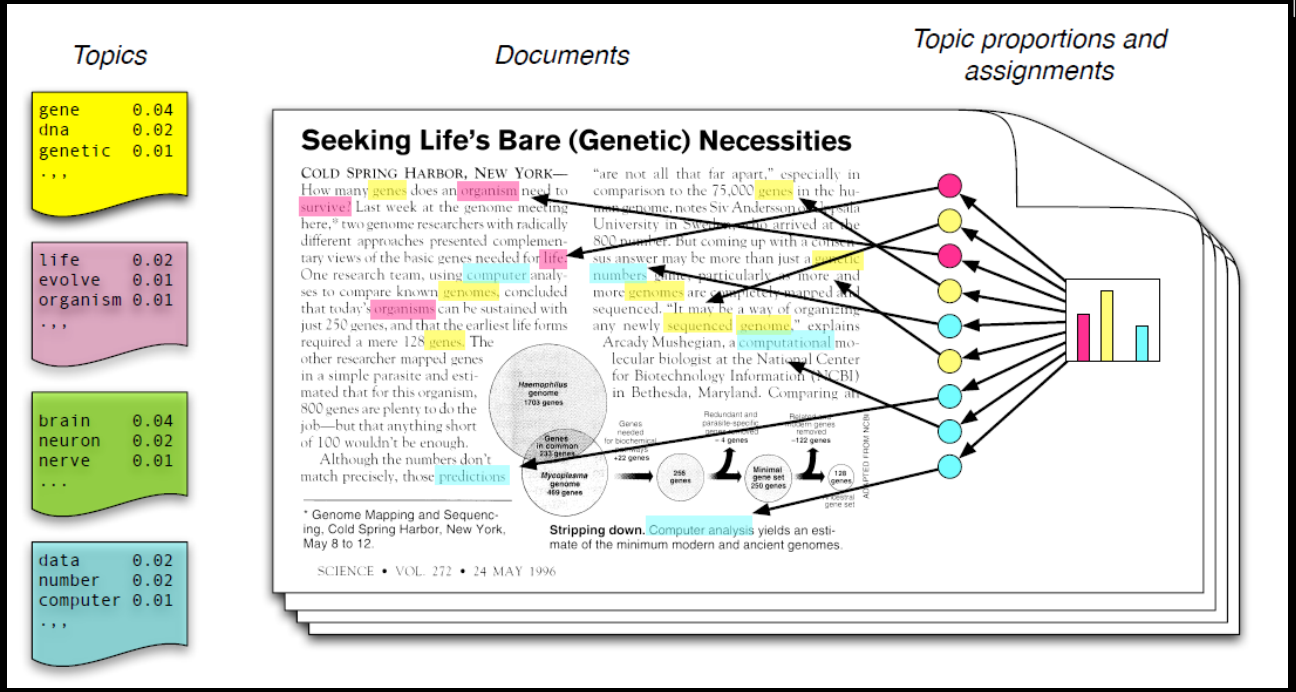
Latent Dirichlet Allocation

- Topics are vectors of probabilities of words



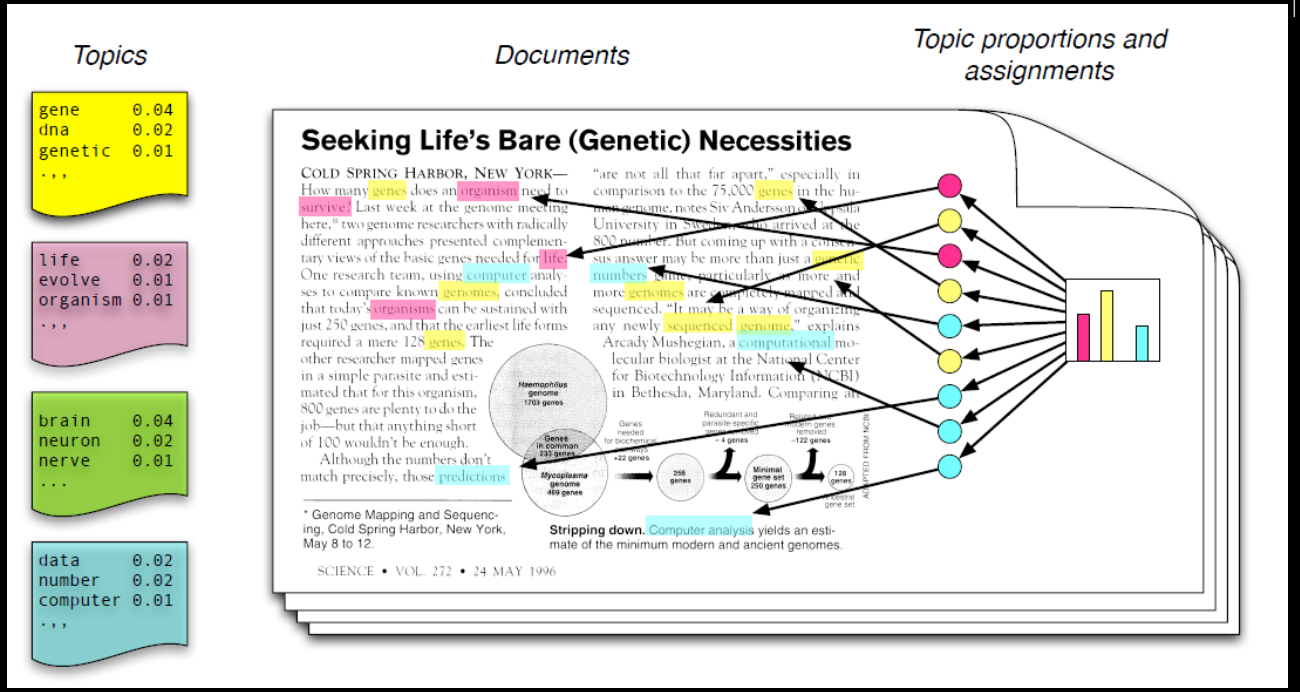
Latent Dirichlet Allocation

- Topics are vectors of probabilities of words
- A document may be a blend of multiple topics



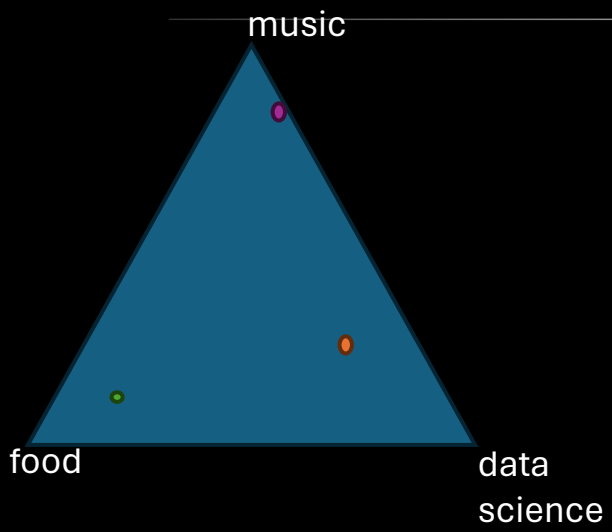
Latent Dirichlet Allocation

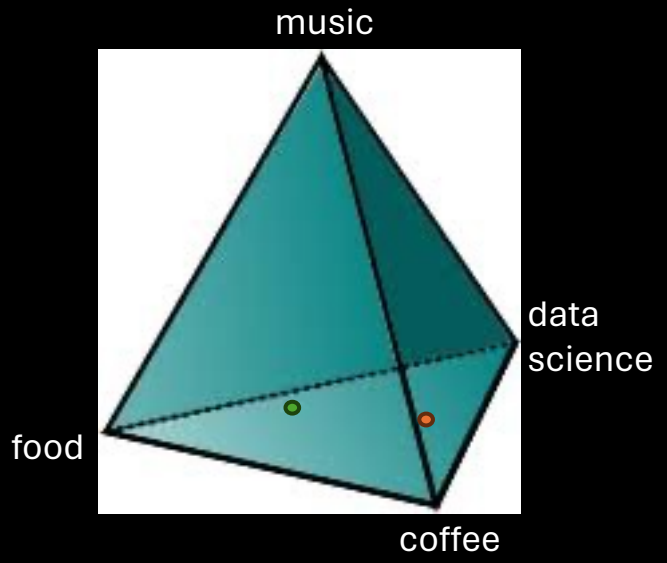
- Topics are vectors of probabilities of words
- A document may be a blend of multiple topics
- The model makes use of *Dirichlet Distributions* to model the relations from Topics to Documents and Words to Topics…



Latent Dirichlet Allocation

- A Dirichlet distribution a generative model...





Bag of Words

- We simplify the data by throwing out word order and just considering the number of occurrences of each word in a document

Bag of Words

- We simplify the data by throwing out word order and just considering the number of occurrences of each word in a document
- We also ignore stop-words like ‘and’, ‘is’, etc

Bag of Words

- We simplify the data by throwing out word order and just considering the number of occurrences of each word in a document
- We also ignore stop-words like ‘and’, ‘is’, etc
- Also numbers, links, punctuation, etc

Bag of Words

- We simplify the data by throwing out word order and just considering the number of occurrences of each word in a document
- We also ignore stop-words like ‘and’, ‘is’, etc
- Also numbers, links, punctuation, etc
- Also anything else we think is going to be frequent across all documents – like ‘UN’ in a corpus of UN tweets

DEMO

Number of topics (k)

- How do we know how many topics we need?

Number of topics (k)

- How do we know how many topics we need?
- Try different numbers of topics and see what gives a sensible result!

Number of topics (k)

- How do we know how many topics we need?
- Try different numbers of topics and see what gives a sensible result!
- We can use coherence scores (a measure of the internal similarity of a topic) to evaluate the quality of a topic or collection of topics

Number of topics (*k*)

- How do we know how many topics we need?
- Try different numbers of topics and see what gives a sensible result!
- We can use coherence scores (a measure of the internal similarity of a topic) to evaluate the quality of a topic or collection of topics
- “It is based on the topic's top 15 words and shows how strongly pairs of these top 15 words support each other within the corpus” (Syed & Spruit 2017)

Number of topics (*k*)

- How do we know how many topics we need?
- Try different numbers of topics and see what gives a sensible result!
- We can use coherence scores (a measure of the internal similarity of a topic) to evaluate the quality of a topic or collection of topics
- “It is based on the topic's top 15 words and shows how strongly pairs of these top 15 words support each other within the corpus” (Syed & Spruit 2017)
- Agrees well with human evaluations of topic quality

DEMO

Compare to BERTopic…

- More pre-processing
- Uses only corpus data
- Instability
- Complexity

DEMO

Thanks Everyone!

Next step: Notebook 3, BYO Data with BERTopic

Will be released by the 12th

Complete in your own time, optional

Please message me on Teams for office hours to discuss!