Modeling Knowledge Incorporation into Topic Models and their Evaluation

Silvia Terragni
s.terragni4@campus.unimib.it
silviatti.github.io
@TerragniSilvia

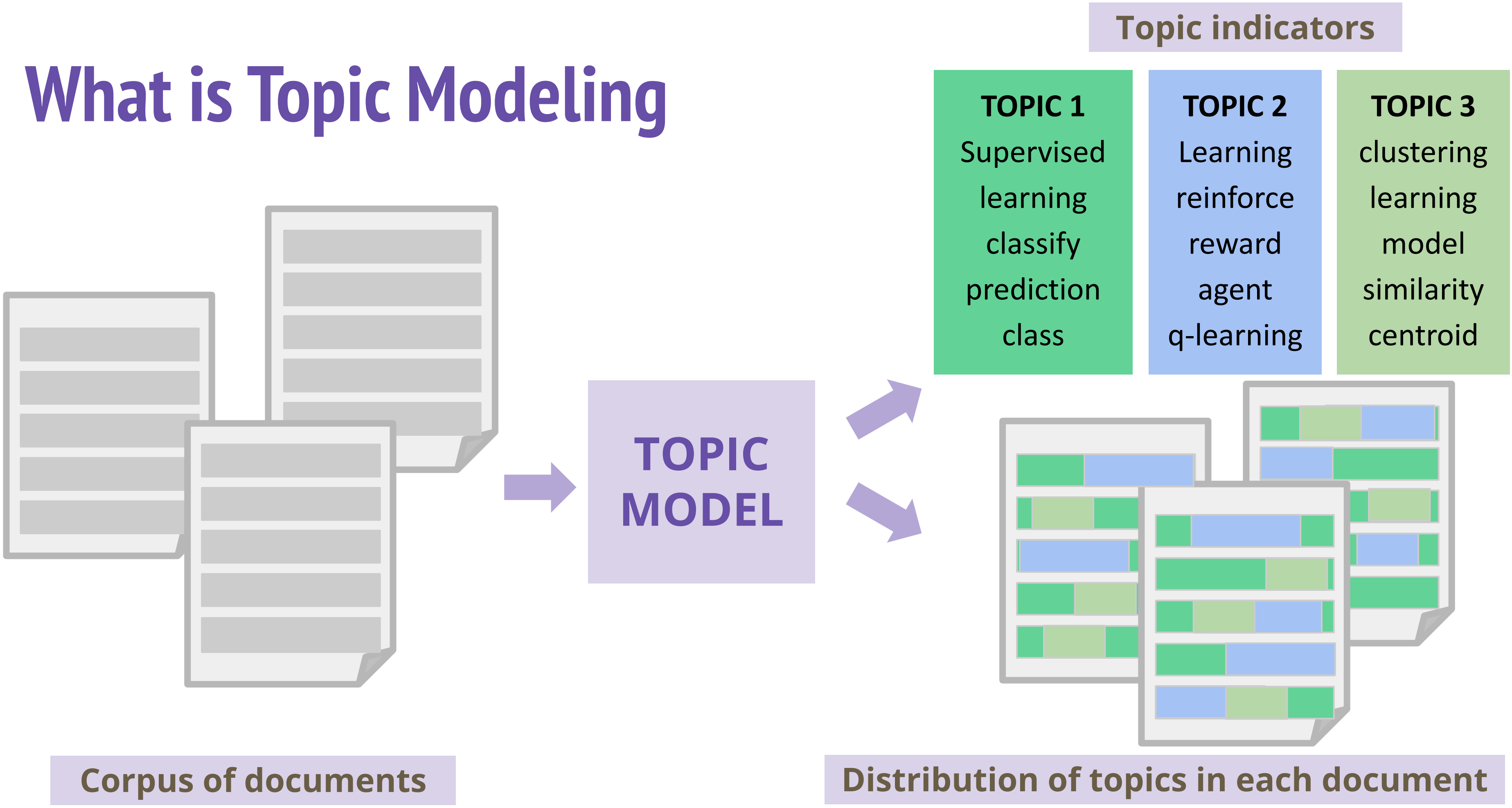
EURECOM (from Milan), 17/06/2021

Outline

- Introduction and state of the art of topic models
- Incorporating knowledge into topic models
  - relationships between documents and words
  - pre-trained contextualized representations
- Evaluation of topic models
  - framework for comparing topic models
  - hyperparameter optimization

Topic Modeling

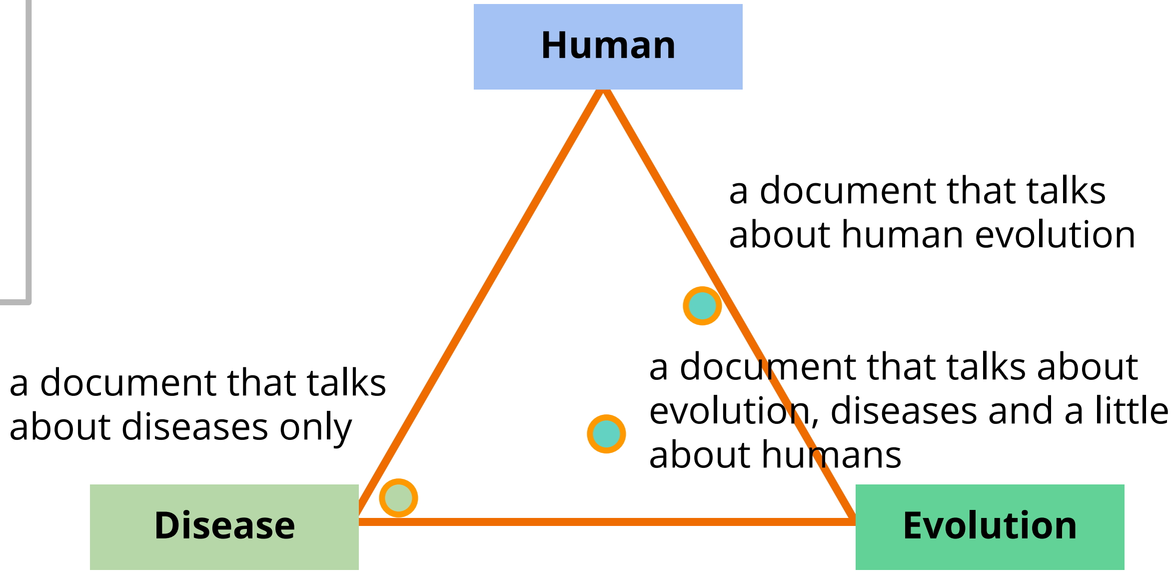
What is Topic Modeling



Topic Models as probabilistic models

The human genome is the complete set of nucleic acid sequences for humans, encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria…

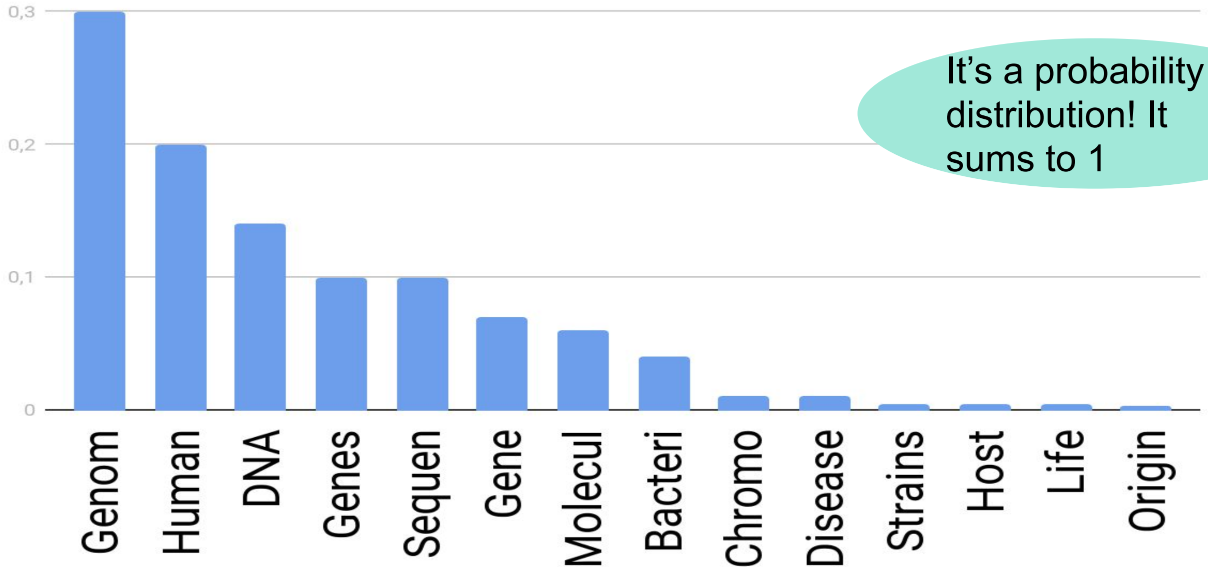
We can express a document as a **multinomial distribution over the topics:** a document talks about different topics in different proportions



Topic Models as probabilistic models

This is not just a unordered list of words. We can expressed it as a **multinomial distribution over the vocabulary**

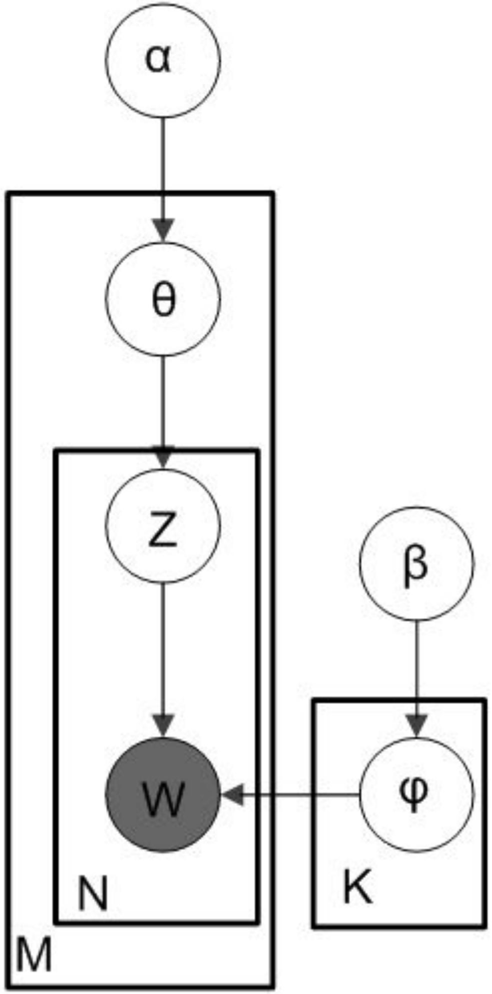
Human
Genome
Dna
Genetic
Genes
Sequence
Gene
Molecular
Map



It's a probability distribution! It sums to 1

Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

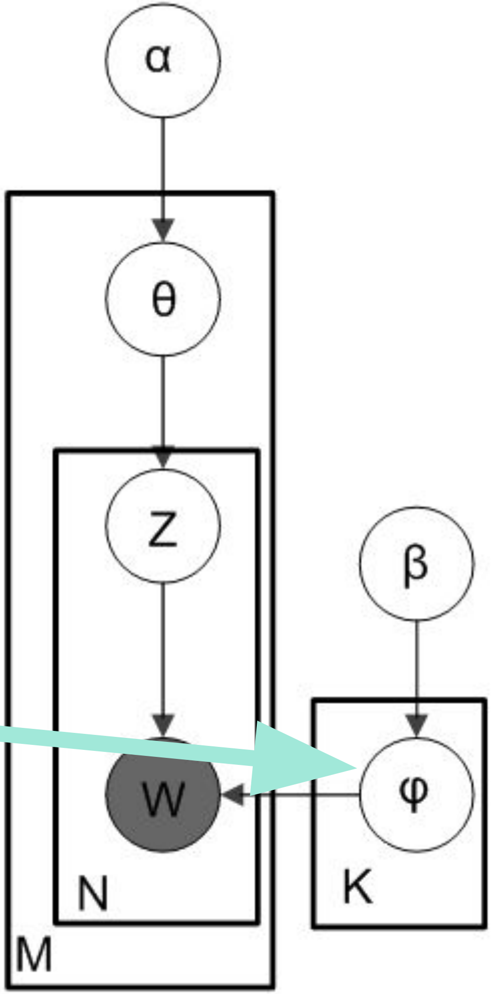


Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

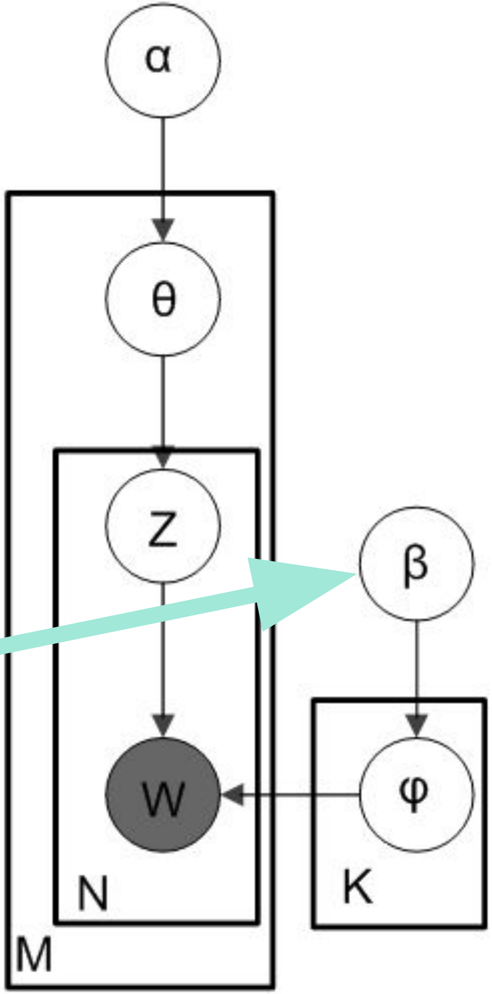
vocabulary

Topics are expressed by a multinomial distribution over the vocabulary



Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

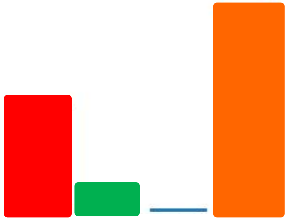


Dirichlet hyperparameter that controls how the sparsity of the words characterizing a topic

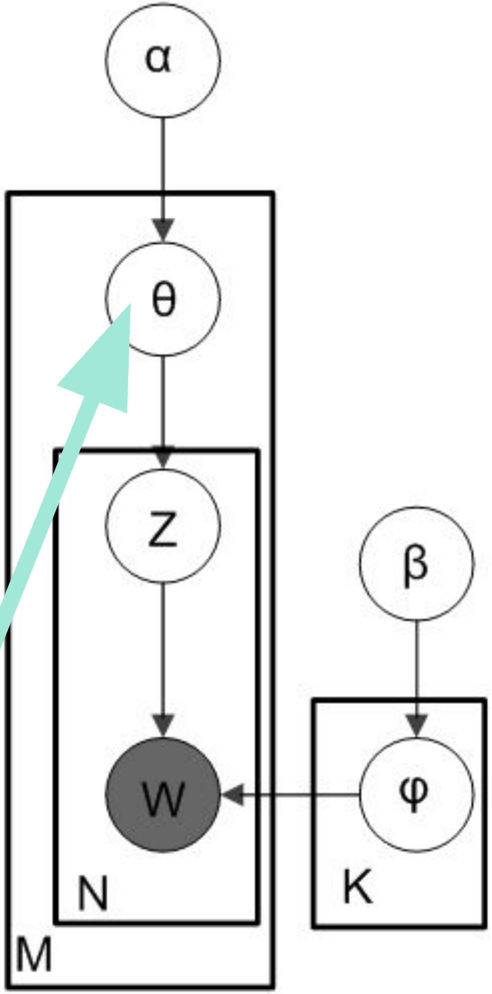
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

Document-topic distribution



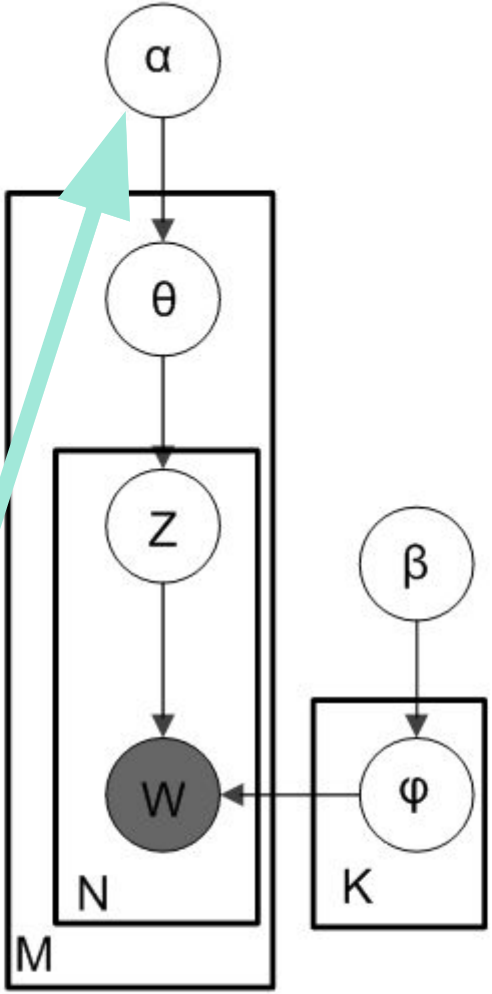
A document is expressed as a multinomial distribution



Latent Dirichlet Allocation

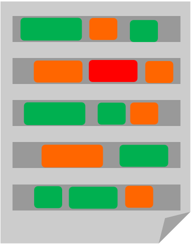
- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

The hyperparameter that controls the sparsity of the topics in a document

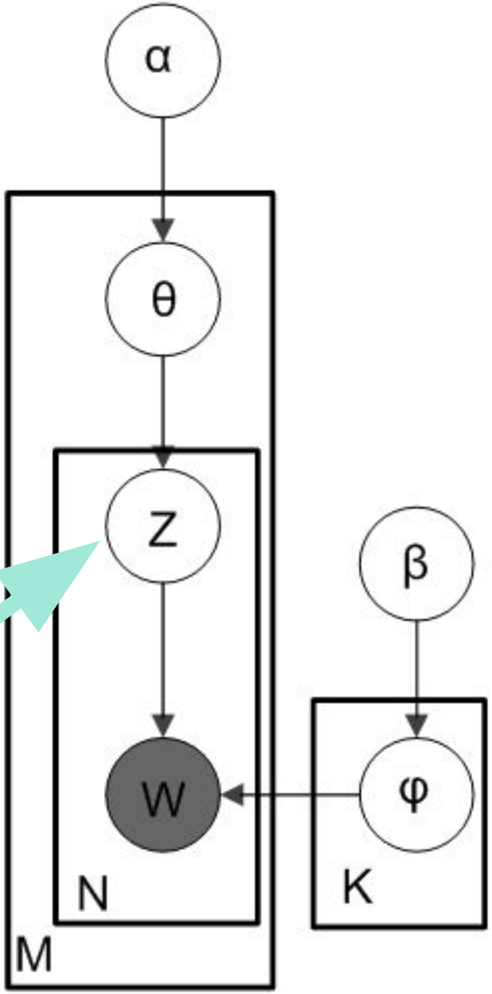


Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)



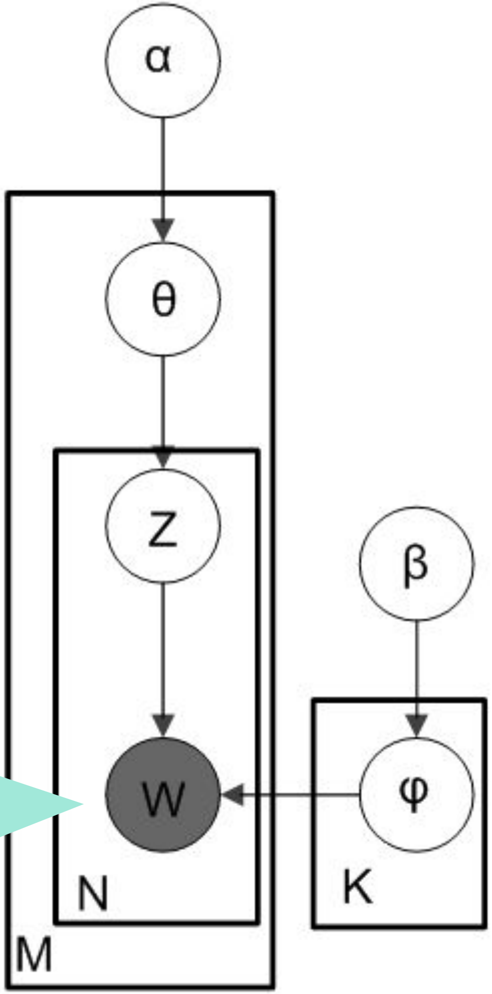
A topic is assigned to each word



Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

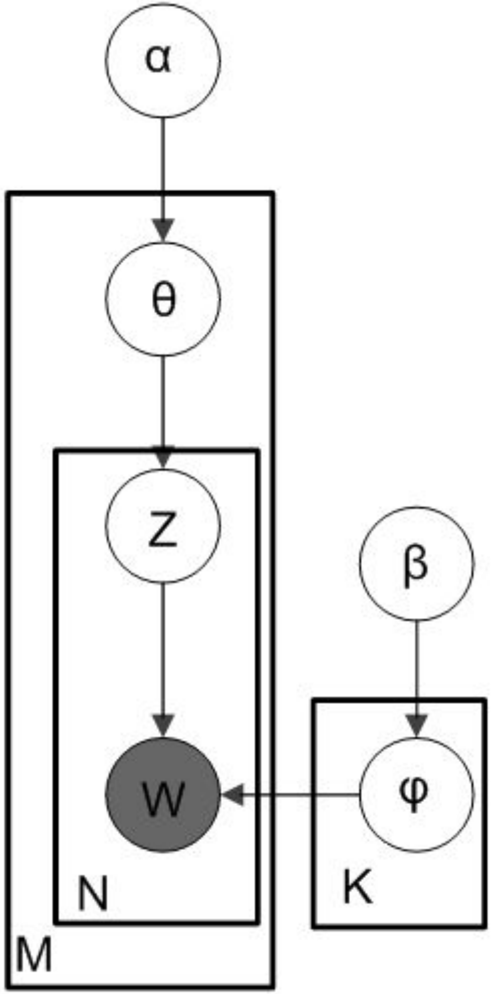
Words are sampled from the word distribution given the topic assignment.



State-of-the-art Topic models

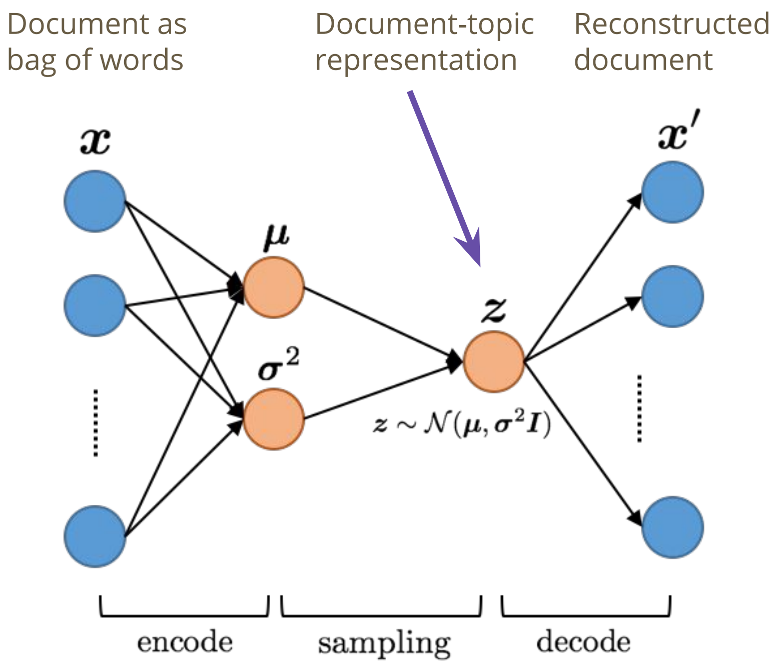
- Usually based on Latent Dirichlet Allocation (LDA) [Blei et al., 2003]

- Increase the capacity of the model by extending LDA:
  - relaxing some assumptions of the model [Wallach et al., 2006]
  - incorporating external knowledge [Nguyen et al., 2015]
  - changing the representation of words [Das et al., 2015]



State-of-the-art Topic models

- Neural Topic Models:
  - usually based on Variational Autoencoders (VAEs) [Miao et al., 2016]
  - the encoder discovers the latent **topic document representation**
  - the **top-words** of a topic are obtained by the weight matrix that reconstructs the BOW



Research Questions

RQ1: **How can we incorporate knowledge into topic models?**

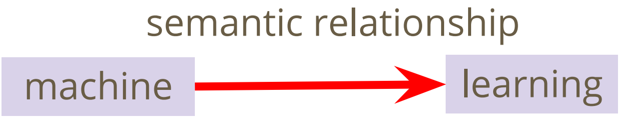
RQ2: **How can we ensure fairer comparisons between the models?**

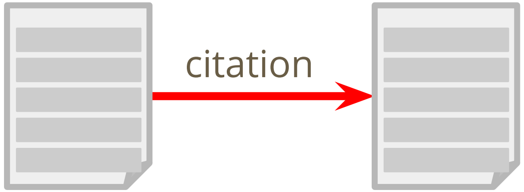
Incorporating Knowledge in Topic Models:

Relationships

Relational Topic Models

Most topic models assume that documents and its constituents (i.e. words) are **independent** from each other





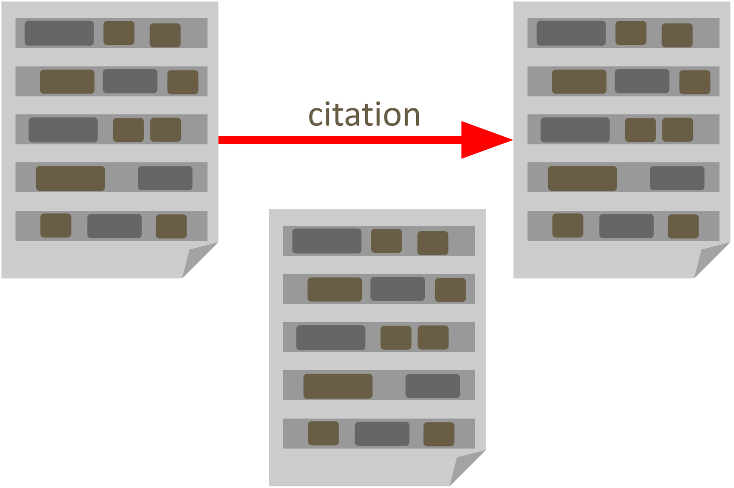
Word-level

[Yang et al, 2015; Nguyen et al, 2015]

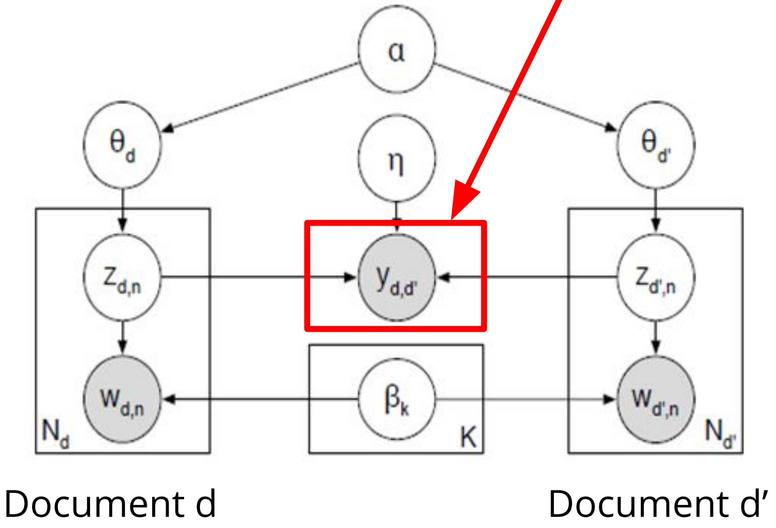
Document-level

[Chang et al., 2009; Yang et al., 2016]

(Document) Relational Topic Models



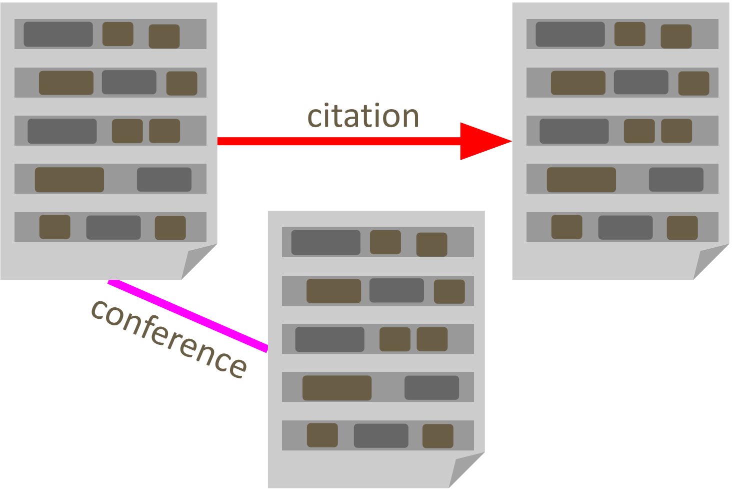
random variable
encodes relationship

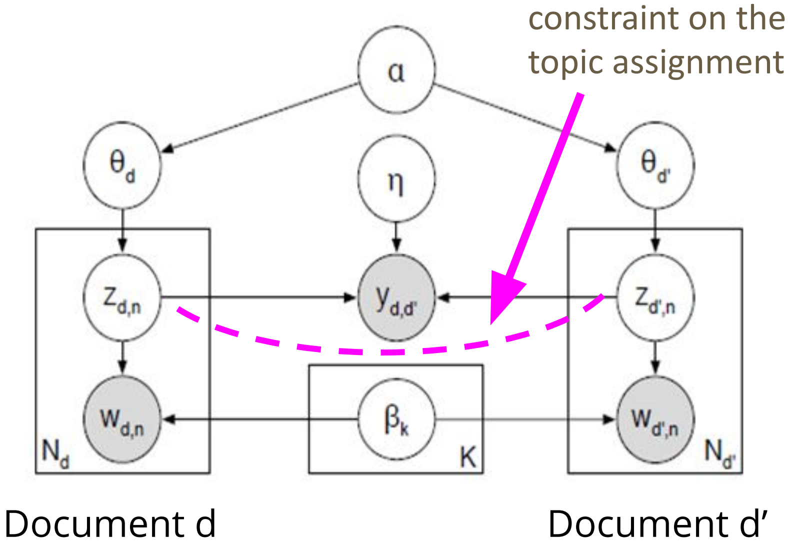


Chang, J.& Blei, D.M.: *Relational Topic Models for Document Networks*. AISTATS 2009: 81-88 (2009)

Document Constrained Relational Topic Models

Document labels in the form of relationships: Two documents that share the same label are more likely to share the same topics

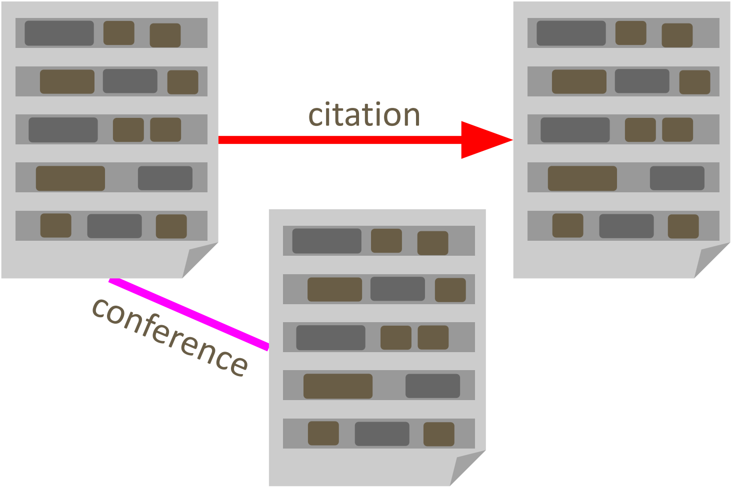


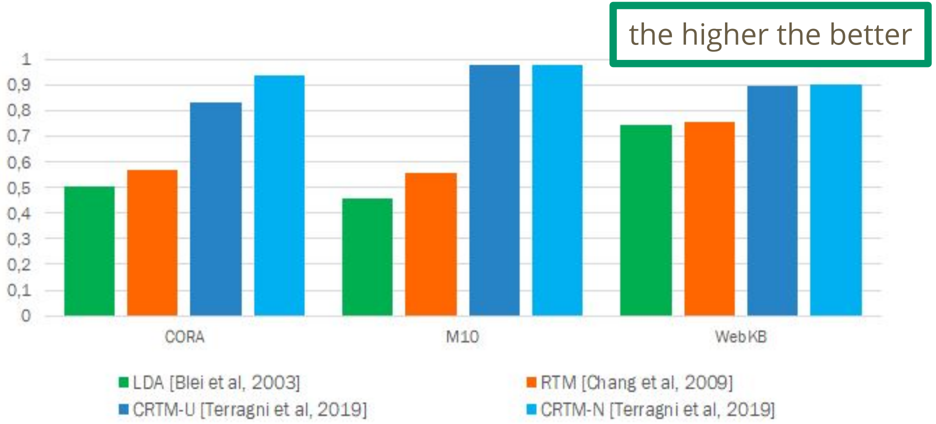


S. **Terragni**, E. Fersini, E. Messina. *Constrained Relational Topic Models.* Information Sciences 512: 581-594 (2020) https://github.com/MIND-Lab/Constrained-RTM

Document Constrained Relational Topic Models

Document labels in the form of relationships: Two documents that share the same label are more likely to share the same topics

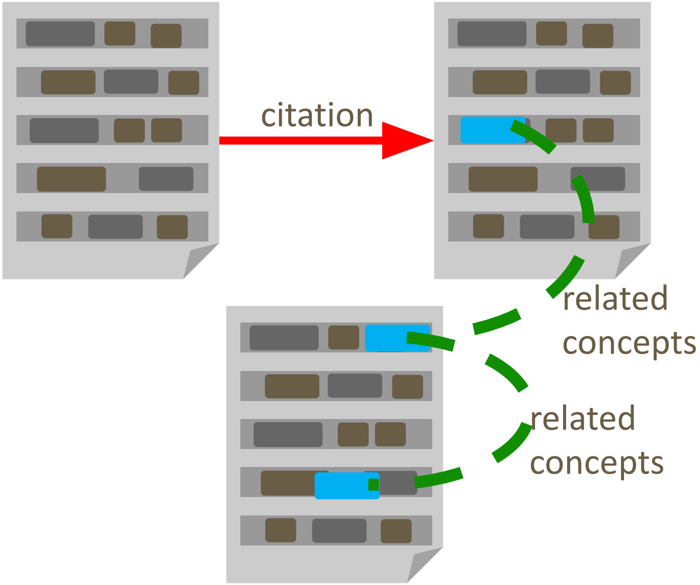




S. Terragni, E. Fersini, E. Messina. *Constrained Relational Topic Models.* Information Sciences 512: 581-594 (2020) https://github.com/MIND-Lab/Constrained-RTM

Entity Constrained Relational Topic Models

- **Relationships between documents** (RTM)
- **Relationships between words and entities:** two named-entities or words that are related are more likely to share the same topics

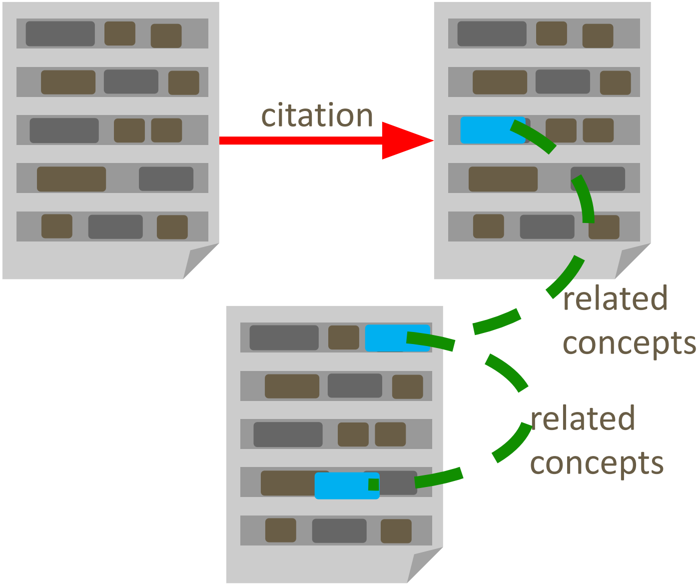


S. Terragni, D. Nozza, E. Fersini, E. Messina. *Which Matters Most? Comparing the Impact of Concept and Document Relationships in Topic Models.* Insights @ EMNLP 2020 [https://github.com/MIND-Lab/EC-RTM]

Entity Constrained Relational Topic Models

- **Relationships between documents** (RTM)
- **Relationships between words and entities:** two named-entities or words that are related are more likely to share the same topics

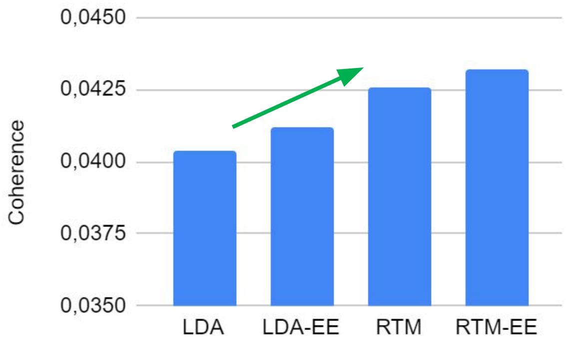
the higher the better



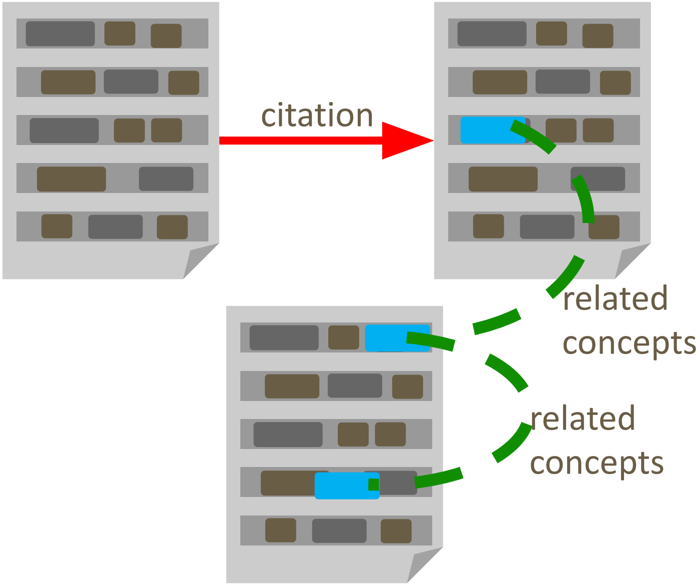
S. Terragni, D. Nozza, E. Fersini, E. Messina. *Which Matters Most? Comparing the Impact of Concept and Document Relationships in Topic Models.* Insights @ EMNLP 2020 [https://github.com/MIND-Lab/EC-RTM]

Entity Constrained Relational Topic Models

- **Relationships between documents** (RTM)
- **Relationships between words and entities:** two named-entities or words that are related are more likely to share the same topics



the higher
the better

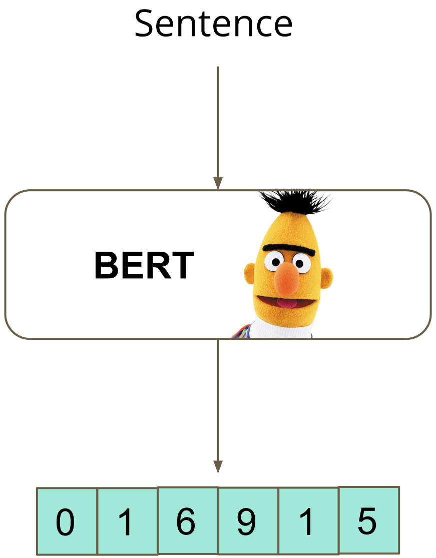


S. Terragni, D. Nozza, E. Fersini, E. Messina. *Which Matters Most? Comparing the Impact of Concept and Document Relationships in Topic Models.* Insights @ EMNLP 2020 [https://github.com/MIND-Lab/EC-RTM]

Incorporating Knowledge in Topic Models:

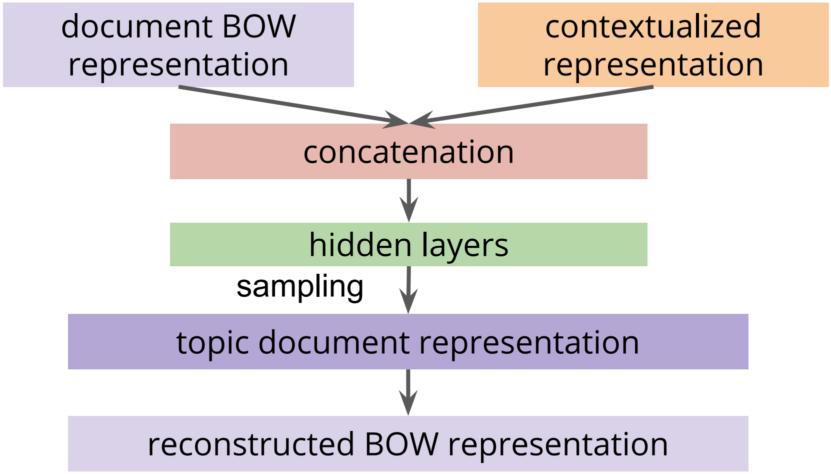
Pre-trained Representations

Why using pre-trained representations



- capture syntactic and semantic information of the sentence
- can be multilingual
- handle out-of-vocabulary (OOV) words

Contextualized Topic Models: Combined CTM



Combined CTM

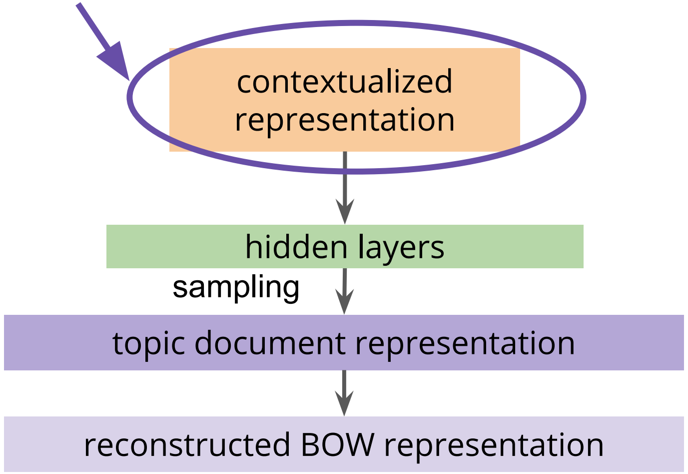
- concatenation of BOW and Sentence BERT
- improve the coherence of the topics
- effective on short texts
- RoBERTa outperforms BERT

Open-source python library: https://github.com/MilaNLProc/contextualized-topic-models
We reached over 32k downloads and 440 github stars :)

Bianchi, F., **Terragni, S.**, & Hovy, D. (2020). *Pre-training is a hot topic: Contextualized document embeddings improve topic coherence*. ACL 2021

Contextualized Topic Models: Zero-shot CTM

also multilingual



What if we replace the BOW representation with pre-trained multilingual representations?

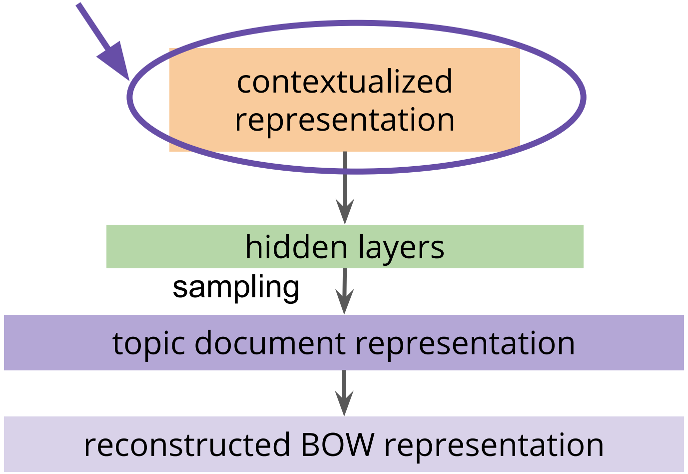
We can **zero-shot predict the topics** of a document in an **unseen language**

Open-source python library: https://github.com/MilaNLProc/contextualized-topic-models
We reached over 32k downloads and 440 github stars :)

Bianchi, F., **Terragni, S.**, Hovy, D., Nozza, D., & Fersini, E. (2020). *Cross-lingual Contextualized Topic Models with Zero-shot Learning.* EACL 2021

Contextualized Topic Models: Zero-shot CTM

also multilingual



Text	Lang	Topic Prediction
Blackmore's Night is a British/American traditional folk....	**EN**	rock, band, bass, formed, ...
I Blackmore's Night sono la band fondatrice del renaissance rock...	**IT**	rock, band, bass, formed, ...
On nomme fourmi de Langton un automate cellulaire...	**FR**	mathematics, theory, space, numbers, ...
Die Ameise ist eine Turingmaschine mit einem zweidimensionalen...	**DE**	mathematics, theory, space, numbers, ...

Bianchi, F., **Terragni, S.**, Hovy, D., Nozza, D., & Fersini, E. (2020). *Cross-lingual Contextualized Topic Models with Zero-shot Learning.* EACL 2021

Evaluating Topic Models

Evaluating a Topic Model

- Evaluating an unsupervised model is not trivial
- Recall that a topic model has two main outputs:

Topic indicators

TOPIC 1	TOPIC 2	TOPIC 3
Supervised	Learning	clustering
learning	reinforce	learning
classify	reward	model
prediction	agent	similarity
class	q-learning	centroid

Topic distribution in each document

Evaluation of the top words

Main aspects of the top words of the topics:

1) how **coherent** are the topics?
2) how **diverse** are the topics?

Evolution	Human	Disease
Evolutionary	Genome	Pizza
Human	Dna	Music
Organisms	Genetic	Diseases
Life	Genes	Sport
Dna	Sequence	Bacterial

Evaluation of the top words

Main aspects of the top words of the topics:

1) how **coherent** are the topics?
2) how **diverse** are the topics?

GOOD TOPICS		JUNK TOPIC
Evolution	Human	Disease
Evolutionary	Genome	Pizza
Human	Dna	Music
Organisms	Genetic	Diseases
Life	Genes	Sport
Dna	Sequence	Bacterial

Some words are not related to others!

Evaluation of the top words

Main aspects of the top words of the topics:

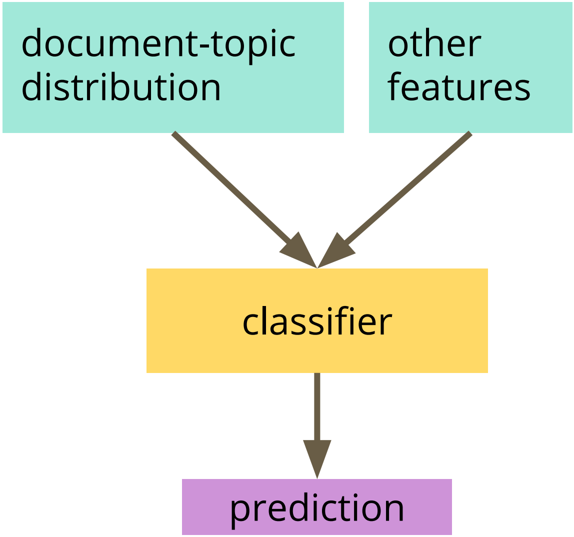
1) how **coherent** are the topics?
2) how **diverse** are the topics?

SIMILAR TOPICS		NOT SIMILAR
Evolution	**Human**	Disease
Evolutionary	Genome	Pizza
Human	**Dna**	Music
Organisms	Genetic	Diseases
Life	Genes	Sport
Dna	Sequence	Bacterial

We'd like that topics express separate ideas or semantic areas

Evaluation of the document-topic distribution

- intrinsic evaluation:
  - **perplexity**: what is the likelihood that the words of the test document x have been generated by the trained topic model?
- extrinsic evaluation:
  - evaluate the **classification** performance
  - any other external task



Why evaluating topic models is hard

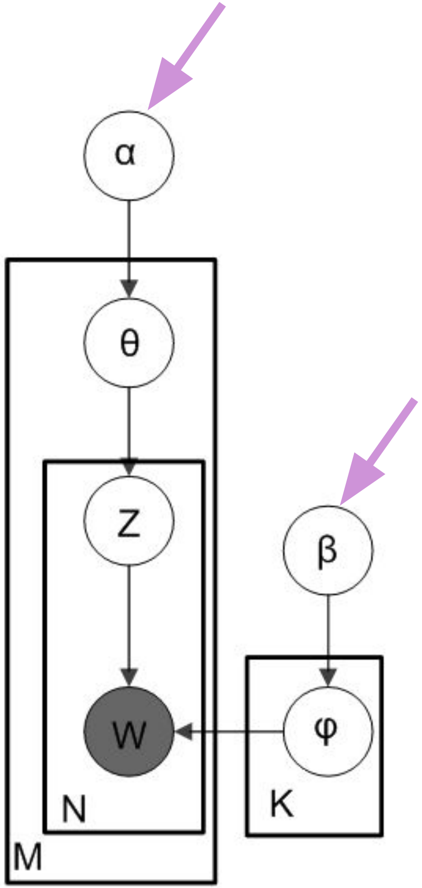
- **No benchmark datasets** and non-standard pre-processing
- **Stochasticity of the results**
- **Which topic model?** Few releases in different programming languages, need to adapt data to each different implementation

A first solution: **ToModAPI**

Lisena, P., Harrando, I., Kandakji, O. & Troncy, R (2020): *TOMODAPI: A Topic Modeling API to Train, Use and Compare Topic Models*, 2nd Workshop for NLP Open Source Software (NLP-OSS)

Why evaluating topic models is hard

- **Hyperparameters setting:**
  - Comparing the models by fixing their hyperparameters is not fair
  - Finding the best hyperparameter configuration is time-consuming



Optimizing and Comparing Topic Models is Simple!

Pre-processing:
- Most common pre-processing tools
- Ready-to-use pre-processed datasets

Evaluation metrics
- Topic coherence
- Topic diversity
- Topic significance
- Document classification

OCTIS

Topic models:
- 4 classical topic models
- 4 neural topic models

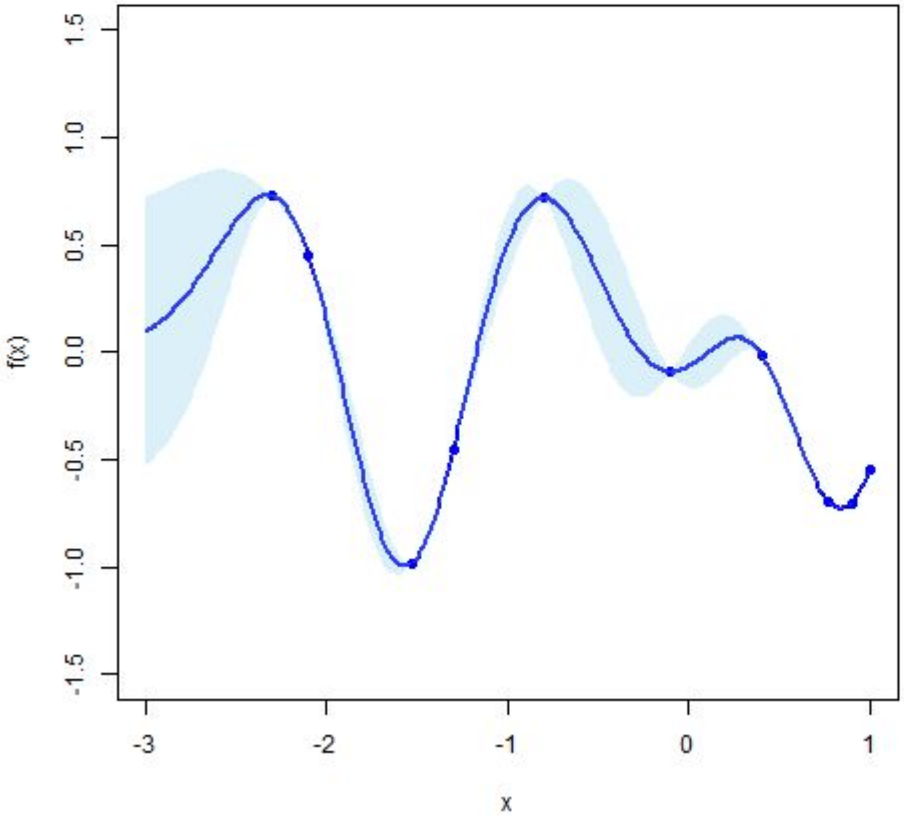
Hyperparameter search
- **Bayesian optimization** for optimizing the hyperparameters

Open-source python library & local web dashboard: https://github.com/mind-lab/octis

We reached over 8k downloads and 170 github stars :)

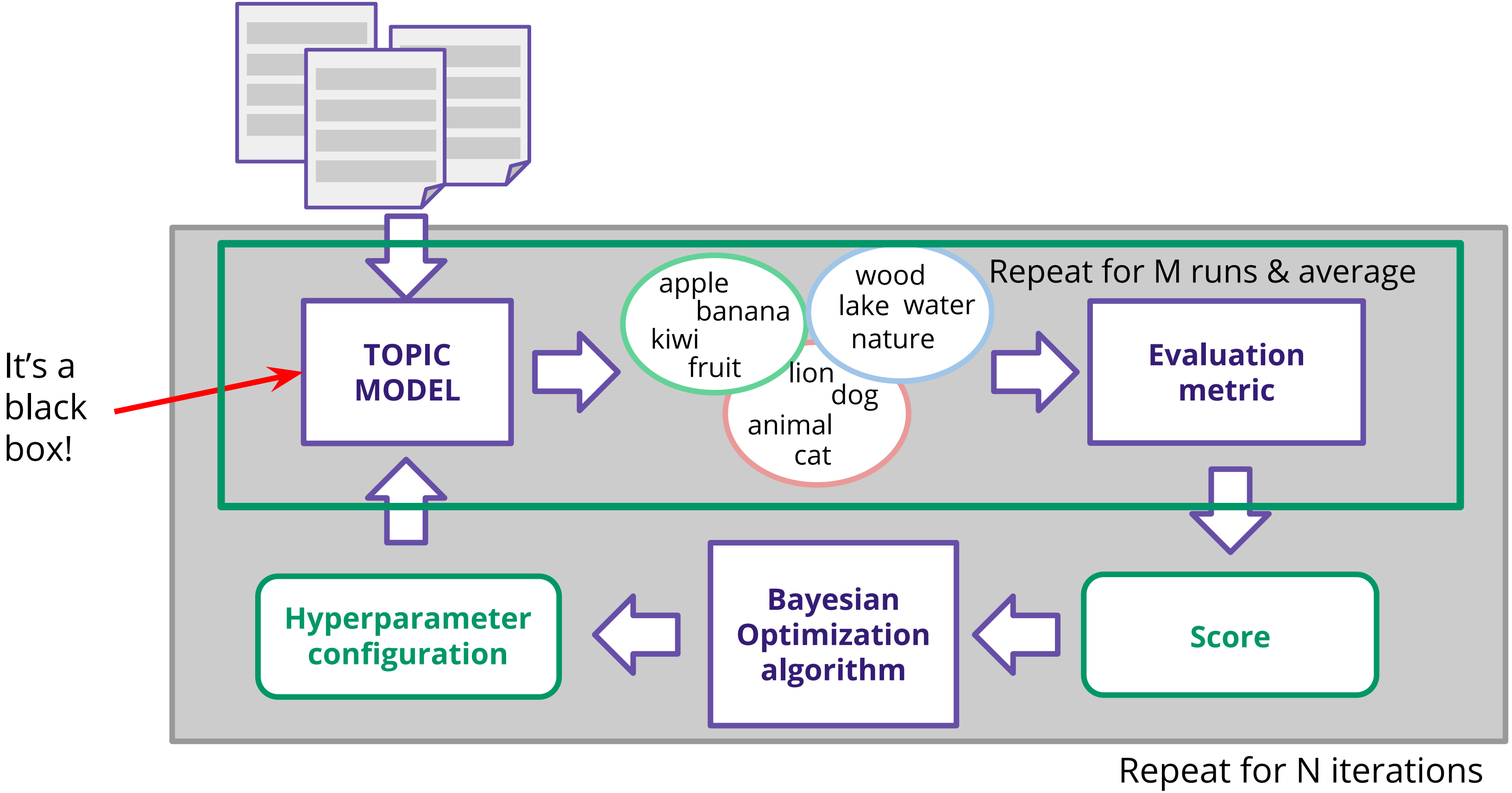
Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). *OCTIS: Comparing and Optimizing Topic models is Simple!*. EACL 2021 (System Demonstrations)

Bayesian Optimization



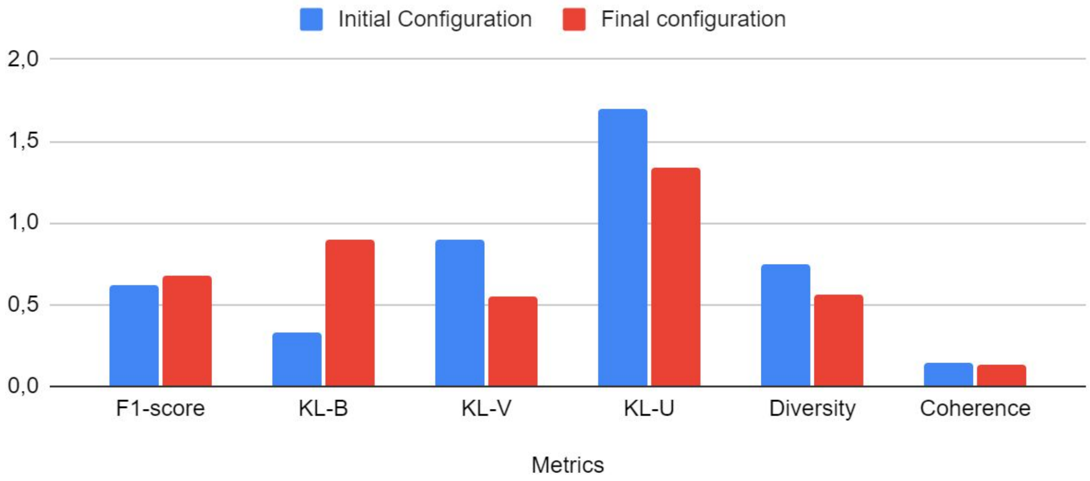
- **probabilistic surrogate model**: approximates the objective function
- **acquisition function**: select the next configuration using the mean and the confidence of the surrogate model

Bayesian Optimization for Topic Modeling



Optimizing the Hyperparameters

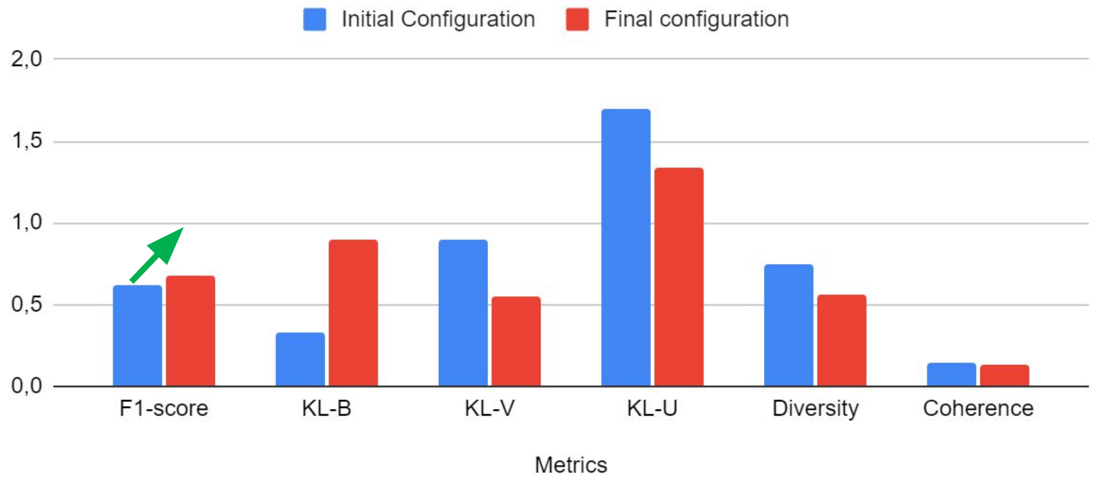
- We optimize the performance of relational topic models with respect to the classification metric F1-score
- We also evaluate other qualitative metrics to investigate different aspects of the RTMs



Optimizing the Hyperparameters

- We optimize the performance of relational topic models with respect to the classification metric F1-score
- We also evaluate other qualitative metrics to investigate different aspects of the RTMs

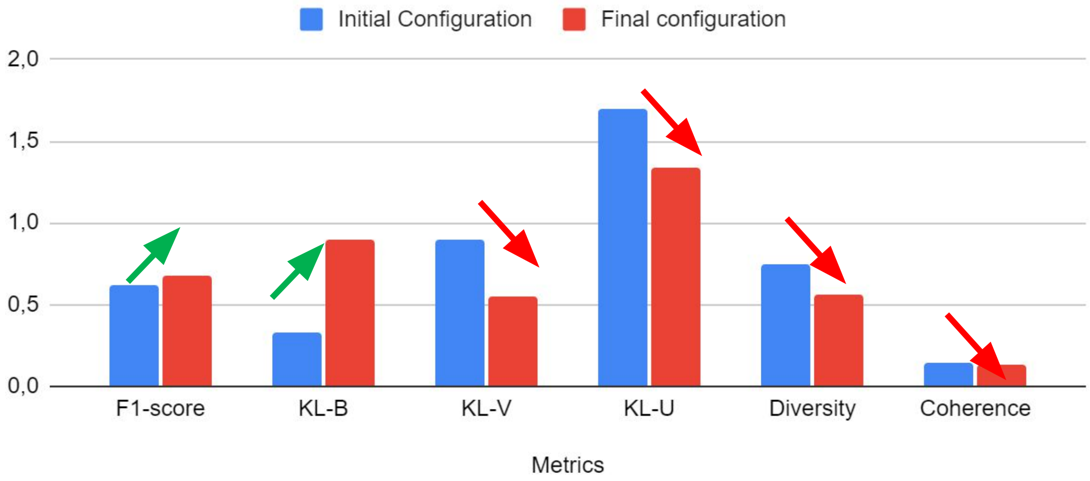
The configuration identified by BO leads to a better performance with respect to its initial configuration



Optimizing the Hyperparameters

- We optimize the performance of relational topic models with respect to the classification metric F1-score
- We also evaluate other qualitative metrics to investigate different aspects of the RTMs

The configuration identified by BO leads to a better performance with respect to its initial configuration



Optimizing for classification purposes can be detrimental to different qualitative metrics

What's next?

- If we optimize for a metric, what happens to the others?

- BO can be expensive:
  - Which hyperparameters are important to optimize?
  - Can we reduce the space of the hyperparameters?
  - Hyperparameter transfer

Thank you :)

References

- Jonathan Chang, David M. Blei: *Relational Topic Models for Document Networks*. AISTATS 2009: 81-88
- Rajarshi Das, Manzil Zaheer, Chris Dyer: *Gaussian LDA for Topic Models with Word Embeddings.* ACL (1) 2015: 795-804
- Yishu Miao, Lei Yu, Phil Blunsom: *Neural Variational Inference for Text Processing*. ICML 2016: 1727-1736
- David M. Mimno, Wei Li, Andrew McCallum: *Mixtures of hierarchical topics with Pachinko allocation.* ICML 2007: 633-640
- Dat Quoc Nguyen, Richard Billingsley, Lan Du, Mark Johnson: *Improving Topic Models with Latent Feature Word Representations*. TACL 3: 299-313 (2015)
- Hanna Wallach: *Topic Modeling: Beyond Bag-of-Words.* ICML 2006
- Yi Yang, Doug Downey, Jordan L. Boyd-Graber: *Efficient Methods for Incorporating Knowledge into Topic Models*. EMNLP 2015: 308-317
- Weiwei Yang, Jordan L. Boyd-Graber, Philip Resnik: *A Discriminative Topic Model using Document Network Structure*. ACL (1) 2016
- Mingyuan Zhou, Yulai Cong, Bo Chen: *Augmentable Gamma Belief Networks*. Journal of Machine Learning Research 17: 163:1-163:44 (2016)