TRAITEMENT AUTOMATIQUE DES LANGUES AVANCE

Master Informatique

2ème Année – 1er Semestre

Intervenants CM:

Gaël DIAS (gael.dias@unicaen.fr), Marc SPANIOL, Fabrice MAUREL, Navneet AGARWAL, Kirill MILINTSEVICH

GREYC Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSICAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

cnrs

Plan de l'UE

1. **[CM 1]** Représentation sémantique de texte [GD]
2. **[CM 2]** Cohérence textuelle [MS]
3. **[CM 3]** Modélisation thématique [NA]
4. **[CM 4]** Résumé de textes et traduction automatique [MS]
5. **[CM 5]** Génération langagière I [KM]
6. **[CM 6]** Génération langagière II [KM]
7. **[CM 7]** TAL multimodal [NA]
8. **[CM 8]** TAL et web [MS]
9. **[CM 9]** TAL et handicap visuel [FM]
10. **[CM 10]** TAL et psychiatrie [GD]

11. **[TP 1-5]** Génération neuronal de comptes-rendus médicaux [NA - KM]

COURS N°1

Modélisation thématique

Plan du cours

- Motivation
- Topic modeling
- Latent Dirichlet Allocation
- Gibbs Sampling
- Nonnegative Matrix Factorization
- Dynamic topic models
- Correlated topic models
- Structured topic models

Motivation

Suppose you are given a massive corpora and asked to carry out following tasks

- Carry out the initial exploratory analysis of the data
- Organise the documents into thematic categories
- Study how these topics evolved over time
- Find relationships between these categories

Topic Models

Topic models are statistical methods that analyze the words within original texts to discover the themes that run through them, study interactions between these themes and also how they evolve over time.

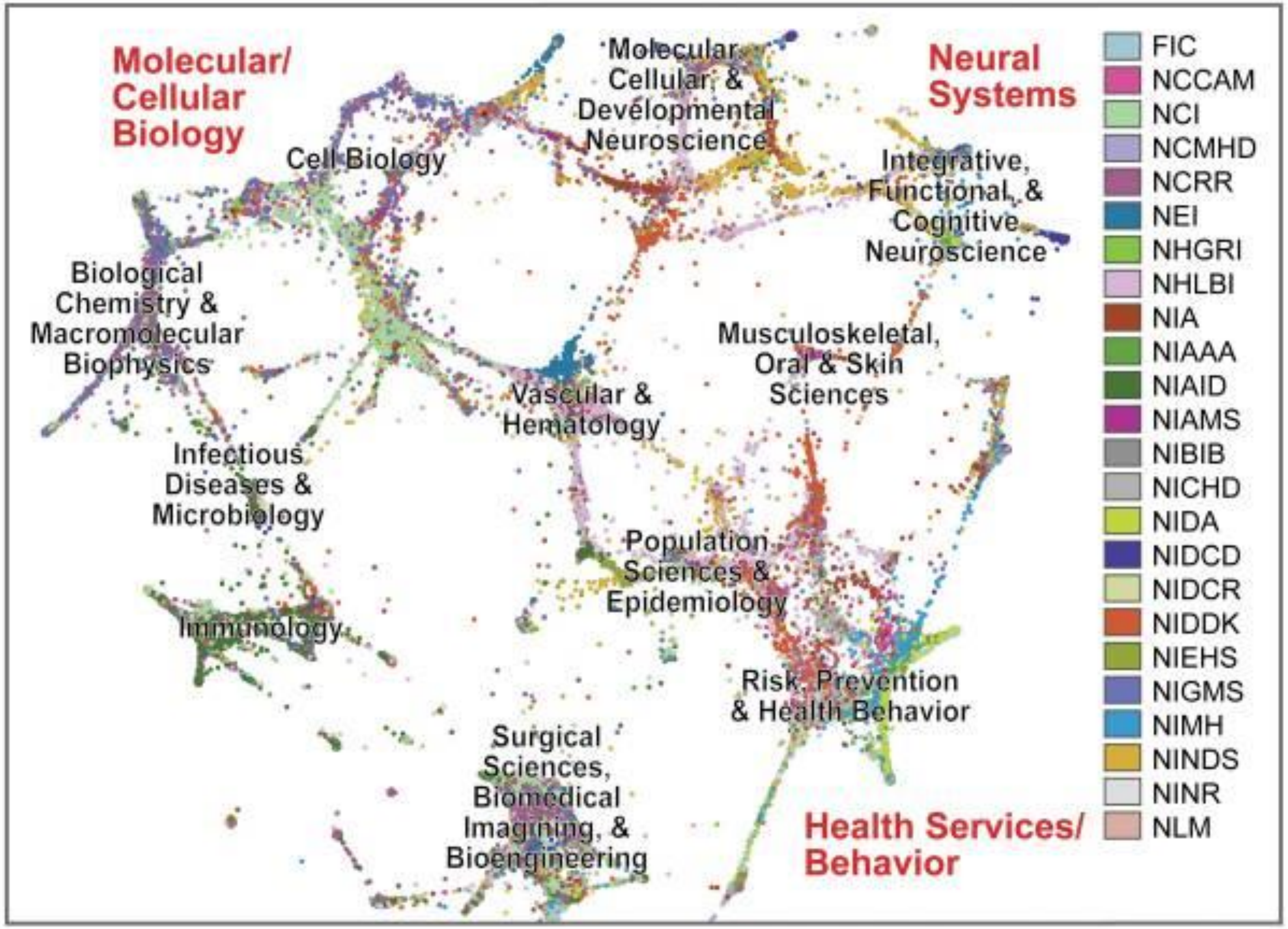
- Unsupervised methods that do not require prior annotations or labeling of documents
- Can be applied to massive collections of documents.
- Applied primarily to text corpora, but concepts are more general
- The topics emerge from the analysis of the original text without the need for human intervention in the learning process.

Topic Models

Map of National Institute of Health grants

year: 2010

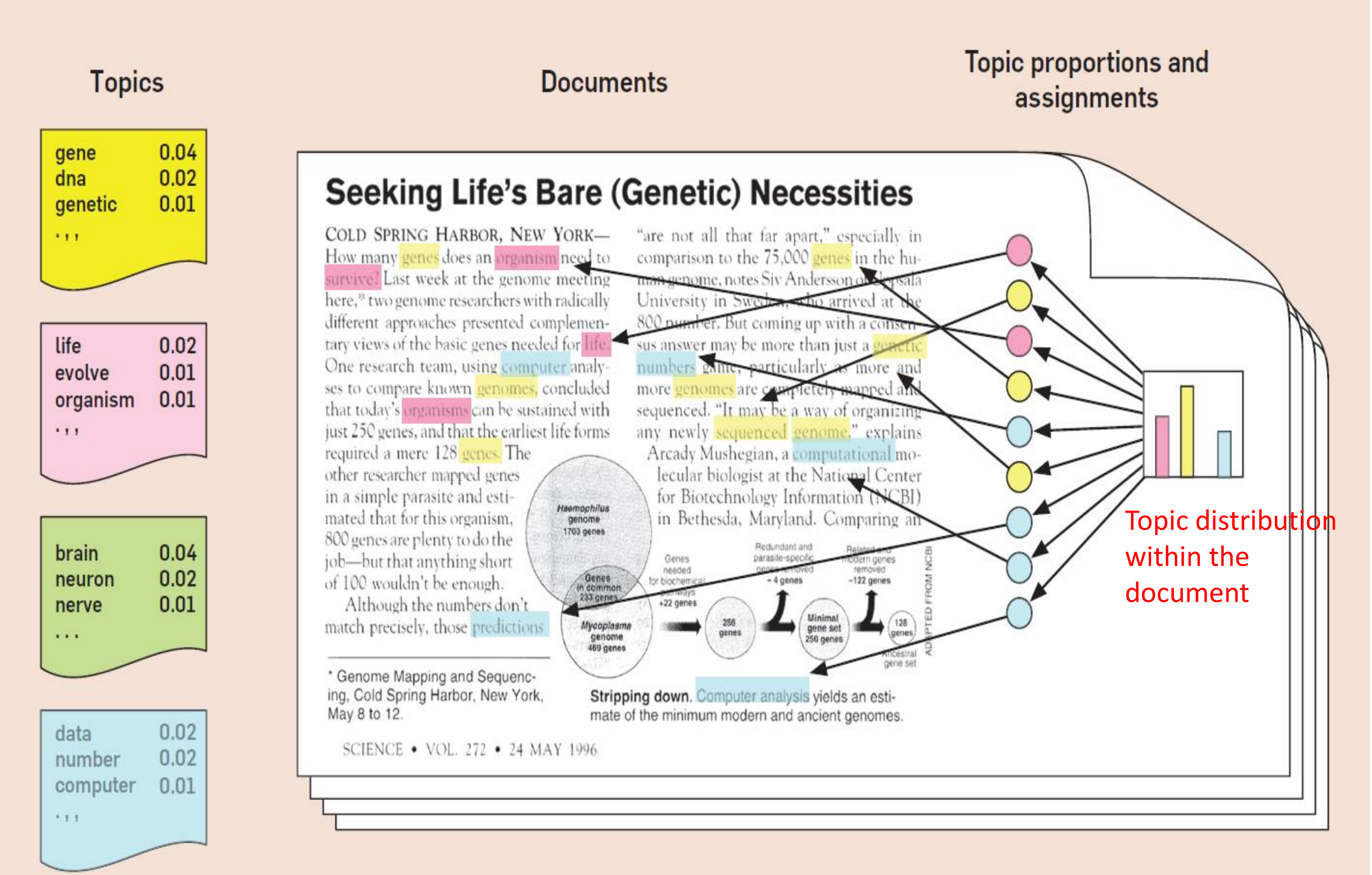
documents: 80,000



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5361216/

Latent Dirichlet Allocation

- Topics are defined to be a distribution over a fixed vocabulary (vocabulary of entire dataset).
- Each document is defined using distributon over topics and each topic is in-turn a distribution over words in the vocabulary
- Topic distribution defines the contribution of each topic towards the document



Blei et al. 2012

Latent Dirichlet Allocation

- Dimensionality reduction:
  - # documents: N
  - # words in vocabulary: V

  documents can be represented using document-term matrix $\mathbb{R}^{N \times V}$

  Assume T topics are learned from this data.

  Documents are represented as distribution over topics $\mathbb{R}^{N \times T}$

	W_1	W_2			W_V
D_1					
D_2					
D_N					

	t_1	t_2			t_T
D_1					
D_2					
D_N					

- Unsupervised learning: can be compared to clustering.

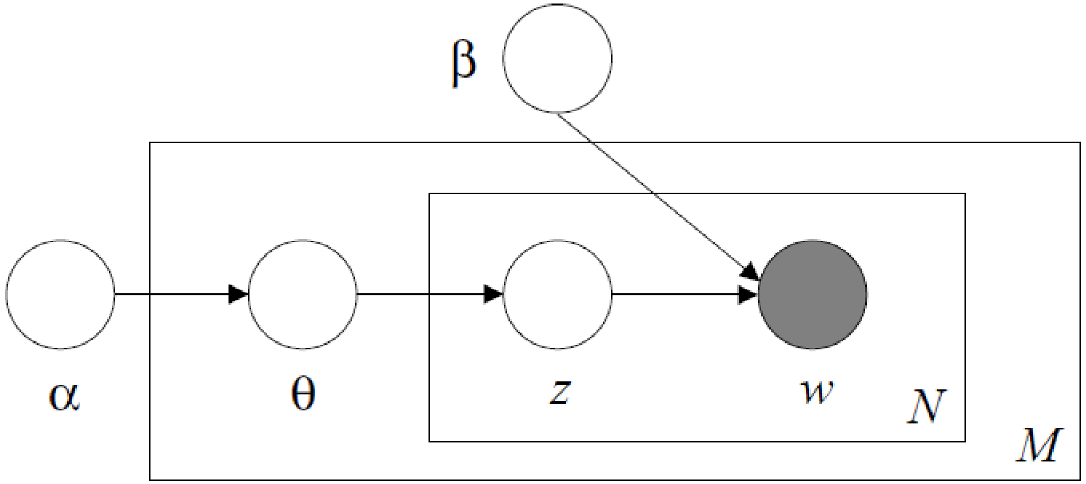
  Words are clustered together to form topics based on their co-occurance patterns

  Documents are clustered based on their topic distributions.

Blei et al., JMLR 2003

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probablistic model of a corpus. Within LDA, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.



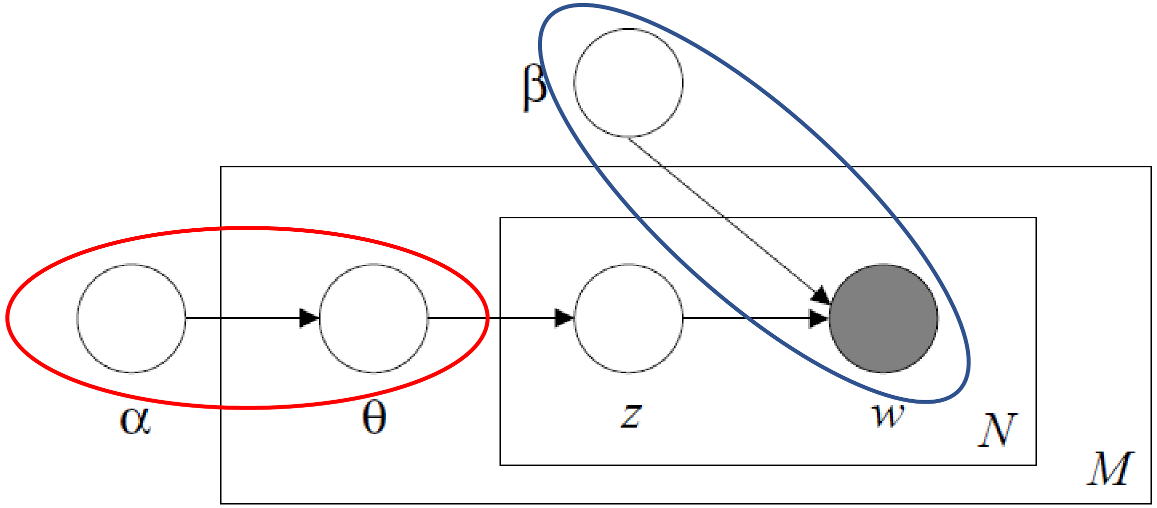
Lets assume we want to generate M documents:

1. Choose N: number of words in the document
2. Choose θ ~Dir(α): topic proportions for document w
3. For each of the N words:
   a) Choose a topic z_n ~ Multinomial(θ): topic assignment for document w
   b) Choose a word w_n from $p(w|z_n, \beta)$

Latent Dirichlet Allocation

Given α and β, the joint distribution of a topic mixture θ, set of *N* topics *z*, and *N* words *w is given by:*

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = \underline{p(\theta \mid \alpha)} \prod_{n=1}^{N} \underline{p(z_n \mid \theta) p(w_n \mid z_n, \beta)},$$



Integrating over θ and summing over *z*, we obtain the marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta.$$

Finally, taking product of marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

Latent Dirichlet Allocation

LDA is part of a larger field of *probabilistic modeling.*

We treat the data as arising from a generative process that includes *hidden variables.*

This generative process defines a joint probability distribution over both the observed and hidden variables

We perform data analysis by using joint distribution to compute conditional distribution of the hidden variables given the observed variables.

This conditional distribution is called *posterior distribution*.

observed variables: *w*

hidden variables: θ, z, β

$$p(\theta, z, \beta | w) = \frac{p(\theta, z, \beta, w)}{p(w)}$$

Latent Dirichlet Allocation

- The posterior cannot be computed because the denominator is intractable.
- Topic modeling algorithms form an approximation of the equation by adapting an alternative distribution over the latent topic structures to be close to the true posterior.
- Topic modeling algorithms generally fall into two categories:
  - Sampling based algorithms
  - Variational algorithms

- The most commonly used sampling algorithm for topic modeling is *Gibbs Sampling*

Gibbs Sampling

- Gibbs sampling procedure considers each word token in the text collection in turn.
- Estimate the probability of assigning the current word token to each topic, conditioned on the topic assignments to all the other word tokens.

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- From this conditional distribution, a topic is sampled and stored as the new assignment for this word token.

Griffiths and Steyvers (2004)

Gibbs Sampling

- n_{dk}: number of words assigned to topic k in document d
- n_{kw}: number of times word w is assigned to topic k
- n_k: number of times any word is assigned to topic k

Input: words $\mathbf{w} \in$ documents $\mathbf{d}$
Output: topic assignments $\mathbf{z}$ and counts $n_{d,k}$, $n_{k,w}$, and n_k
begin
    randomly initialize $\mathbf{z}$ and increment counters
    foreach *iteration* **do**
        for $i = 0 \to N - 1$ **do**
            $word \leftarrow w[i]$
            $topic \leftarrow z[i]$
            $n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1
            for $k = 0 \to K - 1$ **do**
                $p(z = k|\cdot) = (n_{d,k} + \alpha_k)\frac{n_{k,w}+\beta_w}{n_k+\beta\times W}$
            end
            $topic \leftarrow$ sample from $p(z|\cdot)$
            $z[i] \leftarrow topic$
            $n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1
        end
    end
    return $\mathbf{z}$, $n_{d,k}$, $n_{k,w}$, n_k
end

Gibbs Sampling

Example

Assume we have some document with random word-topic assignment

India	enters	world	cup	final
1	3	1	2	4

We have count matrix C^{WT}

	1	2	3	4
India	70	5	0	8
enters	2	3	15	6
world	28	4	12	1
cup	6	43	6	0
final	7	0	9	31

Gibbs Sampling

Example

Assume we have some document with random word-topic assignment

India	enters	world	cup	final
1	3	1	2	4

We have count matrix C^{WT}

	1	2	3	4
India	70	5	0	8
enters	2	3	15	6
world	28	4	12	1
cup	6	43	6	0
final	7	0	9	31

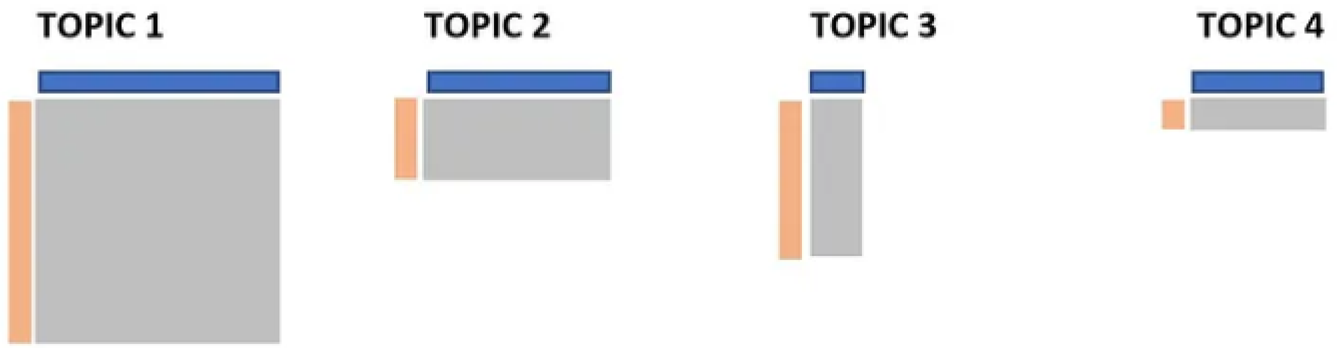
Gibbs Sampling

- *Consider the conribution of each topic towards this document*

- *Next, we take how many times each topic is assigned to this word*

Gibbs Sampling

- *Multiply these values to get conditional probabilities*



- *Finally, pick one of the topics from this distribution and update the variables accordingly.*
- *Repeat this for every word.*

Nonnegative Matrix Factorization (NMF)

1. Nonnegative Matrix Factorization (NMF) factors input nonnegative matrix V into two nonnegative matrices W and H.

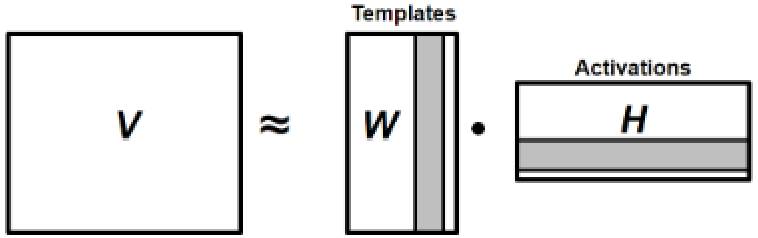


Figure 8.20b from [Müller, FMP, Springer 2015]

V×D V×T T×D

$$V \approx W \cdot H.$$

- W and H are required to have much lower rank that the original matrix V.
- Columns of V contains V-dimensional data vectors
- Columns of W are the word distribution for each topic.
- Rows of H are the activation of given topic across all documents.
- In most cases the factorization does not have an exact solution and requires optimization procedures to find numerical approximations.

$$||V - WH||^2$$

LDA vs NMF

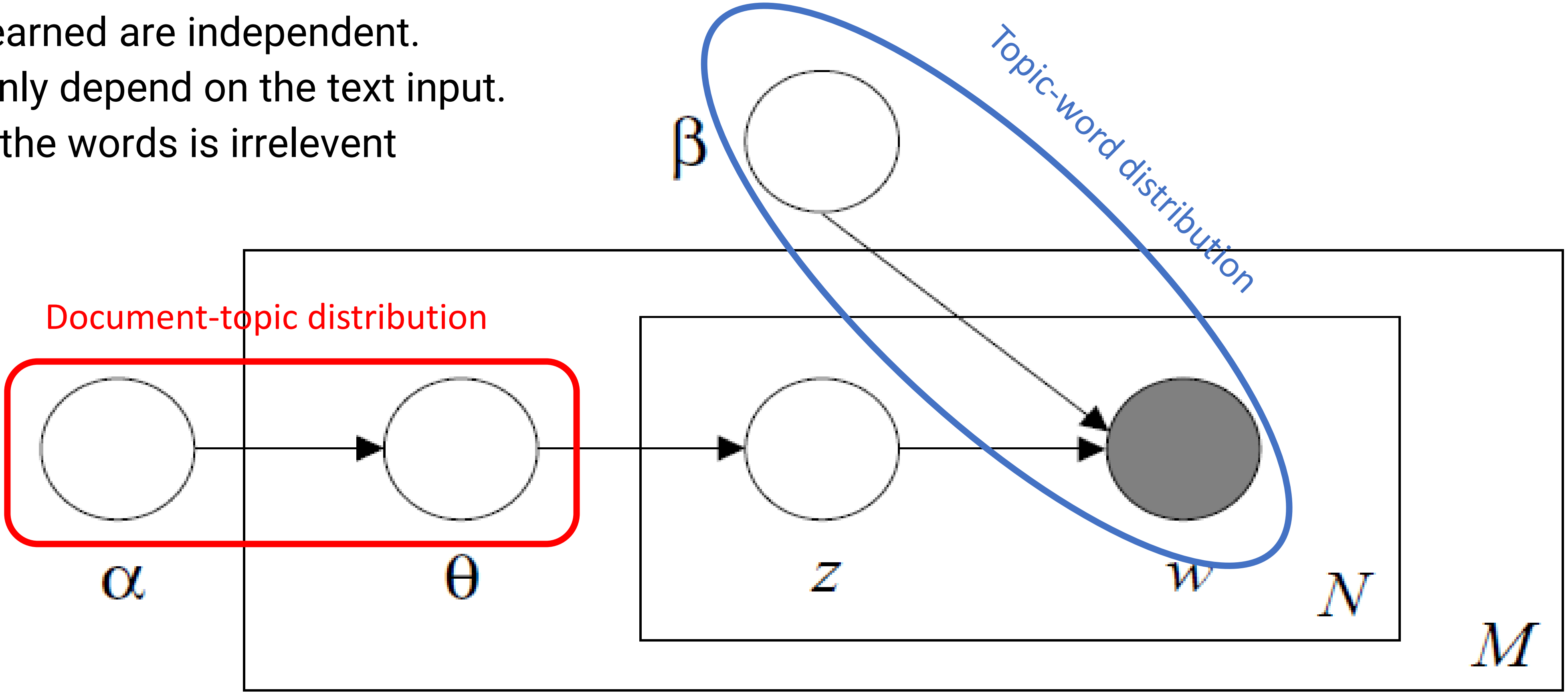
- LDA is probabilistic while NMF uses matrix factorization.
- LDA extracts independent topics from word distributions. Therefore, topics that are dissimilar in the document may not be identified seperately.
- NMF learns dissimilar topics, but can cause difficulties in interpreting findings.
- NMF usually performs better with short texts, like social media data.
- For both LDA and NMF, the results are highly dependent on hyperparameter tuning.

LDA vs NMF

	LDA		NMF	
No.	Topic/content	Keywords	Topic/content	Keywords
1	Government response	ban, travelgov, potus, dv2021, loveisnottourism, whcovidresponse, end, visa, please, vp	Government response	whcovidresponse, potus, loveisnottourism, cdcdirector, presssec, vp, cdctravel, cdcgov, liftthetravelban, cdctravel cdcdirector
2	Association for Molecular Pathology (AMP) / mask and virus	amp, travel, come, spread, mask, place, follow, stay, keep, virus	Association for Molecular Pathology (AMP) / desire to travel	covid, travel, people, amp, want, covid travel, time, travel covid, like, year
3	R_t value / India, UK, Europe	rt, travel, country, India, uk, covid, government, list, eu, news	R_t value	rt, covid, travel, https, covid19, traveler, rt ollysmithtravel, traveler, httpstco, ollysmithtravel
4	Travel restriction / England and Scotland	travel, covid, restriction, city, team, England, despite, event, expect, Scotland	Travel restriction	restriction, travel restriction, covid travel, covid19 travel, ease, covid restriction, travel, lift, covid19 restriction, restriction lift
5	Vaccination / border between Canada and the USA	vaccinate, covid19, international, traveler, travel, vaccination, Canada, border, US, fully	Travel ban / India and UK	ban, India, travel ban, travel India, uk, list, country, ban travel, red, variant
6	Quarantine and lockdown / Australia	traveler, day, quarantine, variant, allow, return, lockdown, Australia, break, two	General about travel / Canada	covid19, travel, covid19 travel, international, travel covid19, country, pandemic, international travel, vaccination, Canada
7	COVID-19 cases / USA	case, new, travel, health, state, tourism, public, number, close, include	Vaccination and quarantine	vaccinate, fully, fully vaccinate, vaccinate covid19, traveler, vaccinate traveler, traveler, quarantine, cdc, require
8	Flight / COVID-19 test	test, travel, need, positive, covid, flight, negative, air, take, airport	COVID-19 cases / New Zealand	case, new, covid case, covid19 case, new case, rise, Zealand, New Zealand, report, case covid19
9	Death / Florida	covid, die, death, cause, florida, child, spike, shoot, traveler002, flu	COVID-19 test	test, covid test, negative, positive, test travel, test positive, PCR, covid19 test, day, result
10	China and USA	travel, covid, call, china, business, 2020, trump, usa, dr	Vaccination pass	vaccine, covid19 vaccine, covid vaccine, passport, vaccine passport, require, vaccine travel, dose, mandate, vaccination
11	Unspecific I	not, covid, vaccine, people, do, travel, get, make, still, would		
12	Unspecific II	travel, may, covid, 2, please, 1, help, show, 3, pass		
13	Unspecific III	covid19, travel, due, pandemic, world, today, first, update, coronavirus, safe		
14	Unspecific IV	covid, be, go, travel, time, get, want, one, year, see		

Topic model variations

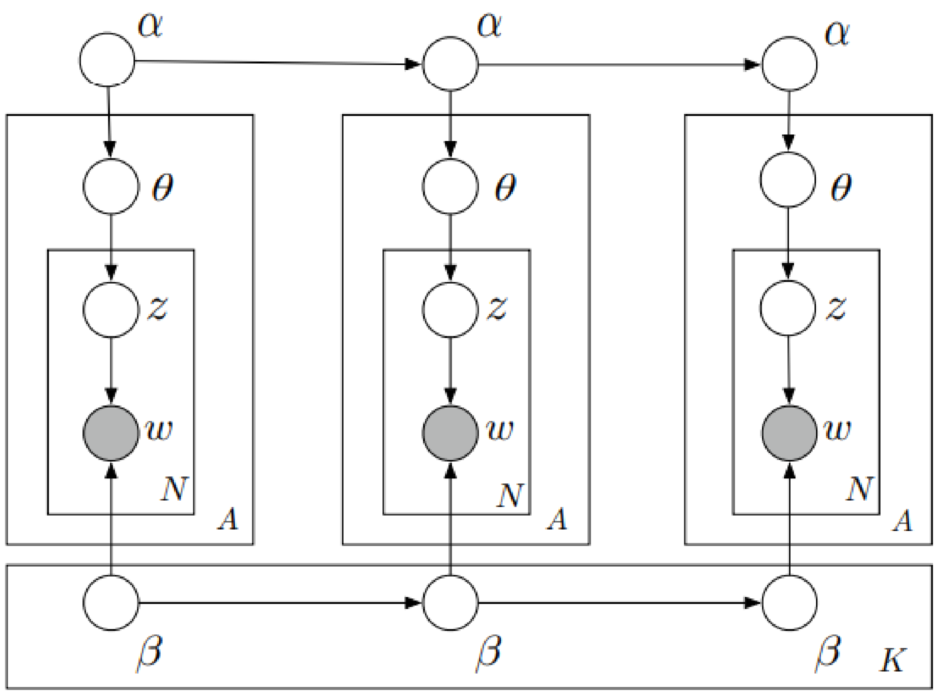
1. Order of documents does not matter.
2. Topics learned are independent.
3. Topics only depend on the text input.
4. Order of the words is irrelevent



Blei, David M., and John D. Lafferty. "Dynamic topic models." *Proceedings of the 23rd international conference on Machine learning*. 2006.

Topic model variations

1. Dynamic topic model
   - LDA assumes that order of documents does not matter.
   - This assumption may be unrealistic when considering long running collections that span years or centuries.
   - Dynamic topic model solves this problem by dividing documents based on time slots.
   - Topic models learned for each time slot are dependent on respective topic from previous time slot.

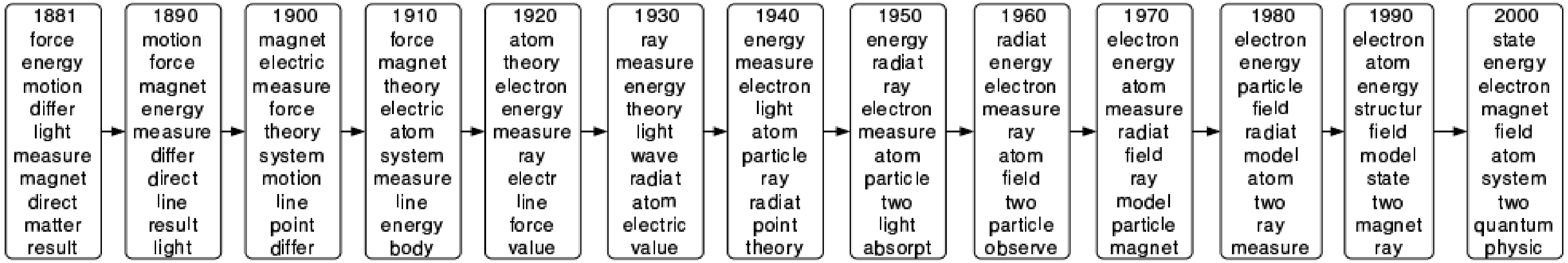


1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim Mult(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

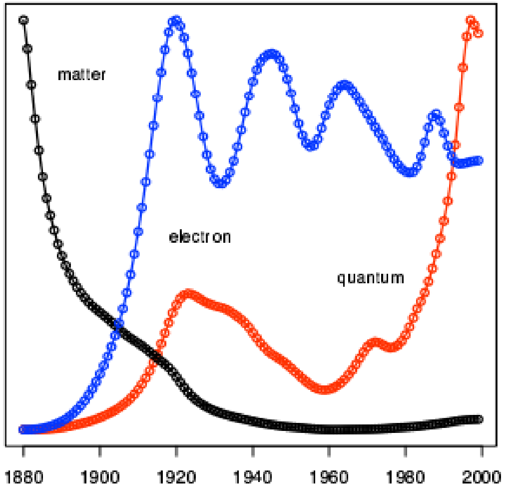
Blei, David M., and John D. Lafferty. "Dynamic topic models." *Proceedings of the 23rd international conference on Machine learning*. 2006.

Topic model variations

1. Dynamic topic model
   - Subset of 30,000 articles from *Science*, 250 from each of the 120 years between 1881 and 1999



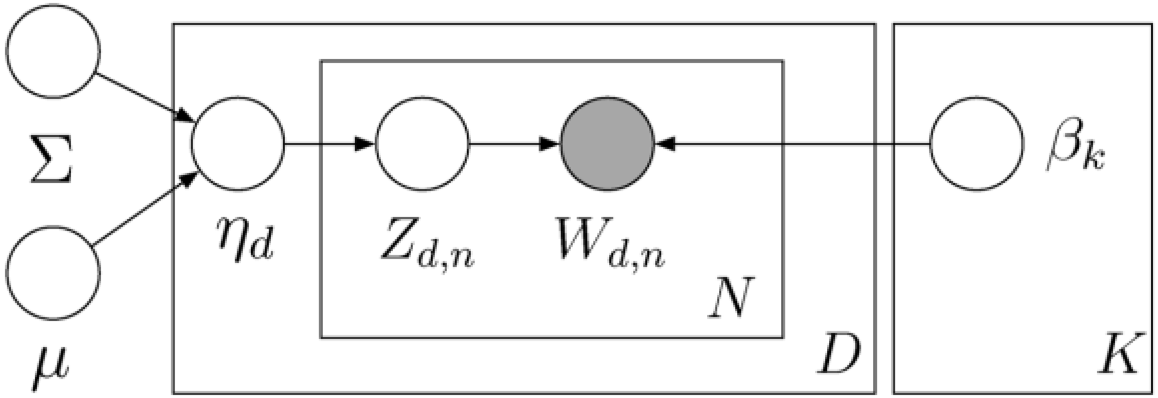
"Atomic Physics"



1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball''
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

Topic model variations

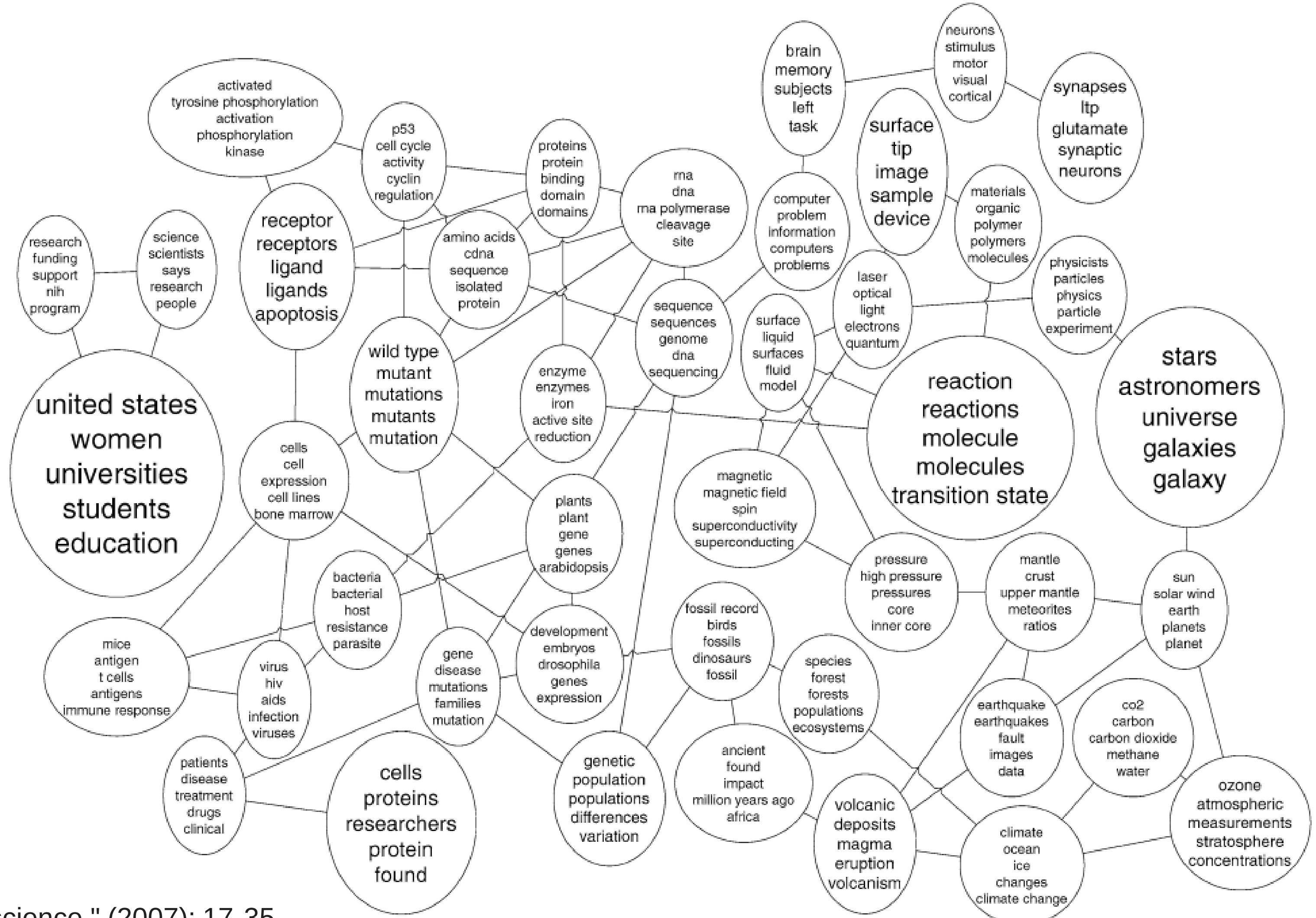
1. Correlated topic model
    - Within LDA, topics are sampled from a Dirichlet distribution are independent which is not realistic for real document collections.
    - CTM draws real values random vectors from multivariate Gaussian distribution inducing dependencies between the components.



Blei, David M., and John D. Lafferty. "A correlated topic model of science." (2007): 17-35.

Topic model variations

1. Correlated topic model
   - Topic graph learned from 16,351 OCR articles from science (1990-1999)



Blei, David M., and John D. Lafferty. "A correlated topic model of science." (2007): 17-35.

Topic model variations

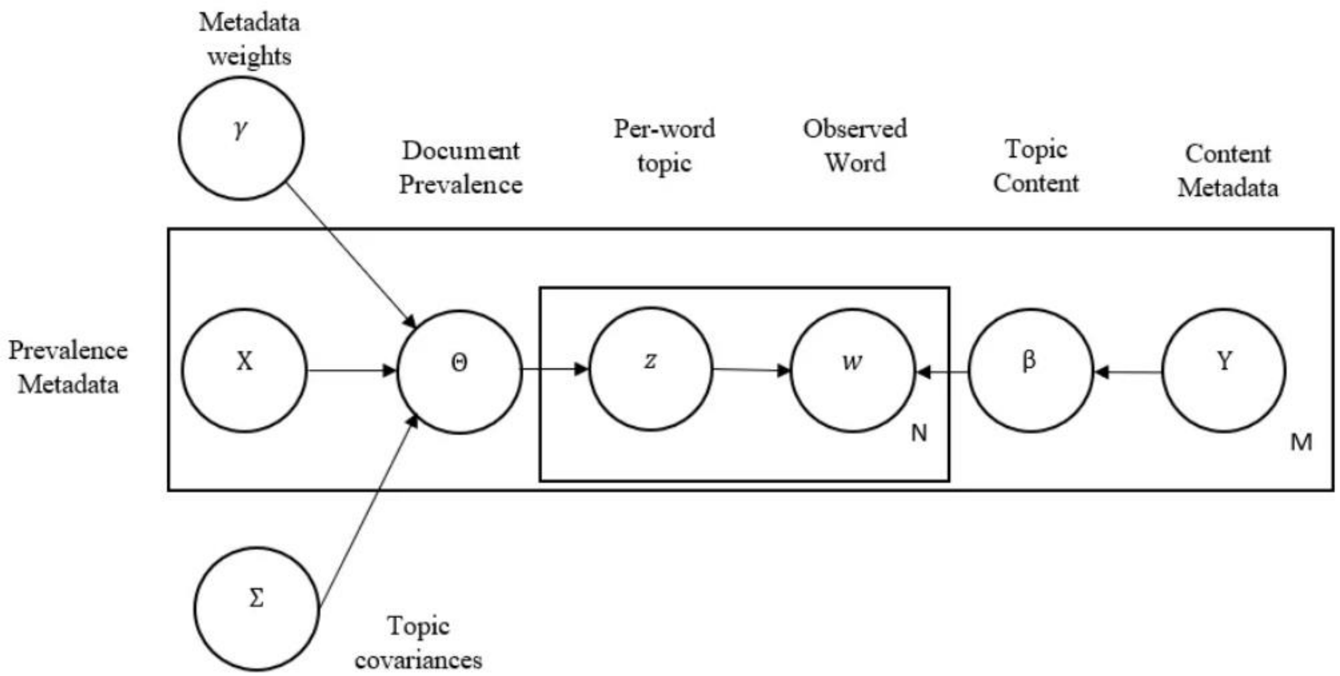
1. Structured Topic Model (STM)

   - LDA learns only based on input text.
   - Certain sources may be more likely to write about politics.
   - Metadata can include date published, author, publication, likes on social media, etc.
   - Within LDA our topic distribution comes from Dirichlet distrubution.
   - STM defines topic distributions based on document metadata
   - We need to go from X_i, *1xp* metadata vector to *1xk* vector of topic distribution.
   - We multiply X with pxk weight matrix t.

Les transformeurs avec récurrence

1. Introduire de la récurrence dans les transformeurs
   - (Dai et al. 2019) proposent de sérialiser le traitement des séquences en gardant en mémoire les valeurs d’activation (valeurs d’attention et couches cachées) du segment précédent.
   - Ce modèle appelé Transformer-XL introduit aussi la notion de plongement de position relative.

$$\theta_i \sim \mathrm{LogisticNormal}(\tau X_i, \Sigma)$$



COURS N°1

Modélisation thématique

Questions supplémentaires?

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Plan de l'UE

1. **[CM 1]** Représentation sémantique de texte [GD]
2. **[CM 2]** Cohérence textuelle [MS]
3. **[CM 3]** Modélisation thématique [NA]
4. **[CM 4]** Résumé de textes et traduction automatique [MS]
5. **[CM 5]** Génération langagière I [KM]
6. **[CM 6]** Génération langagière II [KM]
7. **[CM 7]** TAL multimodal [NA]
8. **[CM 8]** TAL et web [MS]
9. **[CM 9]** TAL et handicap visuel [FM]
10. **[CM 10]** TAL et psychiatrie [GD]

11. **[TP 1-5]** Génération neuronal de comptes-rendus médicaux [NA - KM]

TRAITEMENT AUTOMATIQUE DES LANGUES AVANCE

Master Informatique

$2^{\text{ème}}$ Année – 1^{er} Semestre

Intervenants CM:

Gaël DIAS (gael.dias@unicaen.fr), Marc SPANIOL, Fabrice MAUREL

Navneet AGARWAL, Kirill MILINTSEVICH

GREYC Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSICAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

cnrs