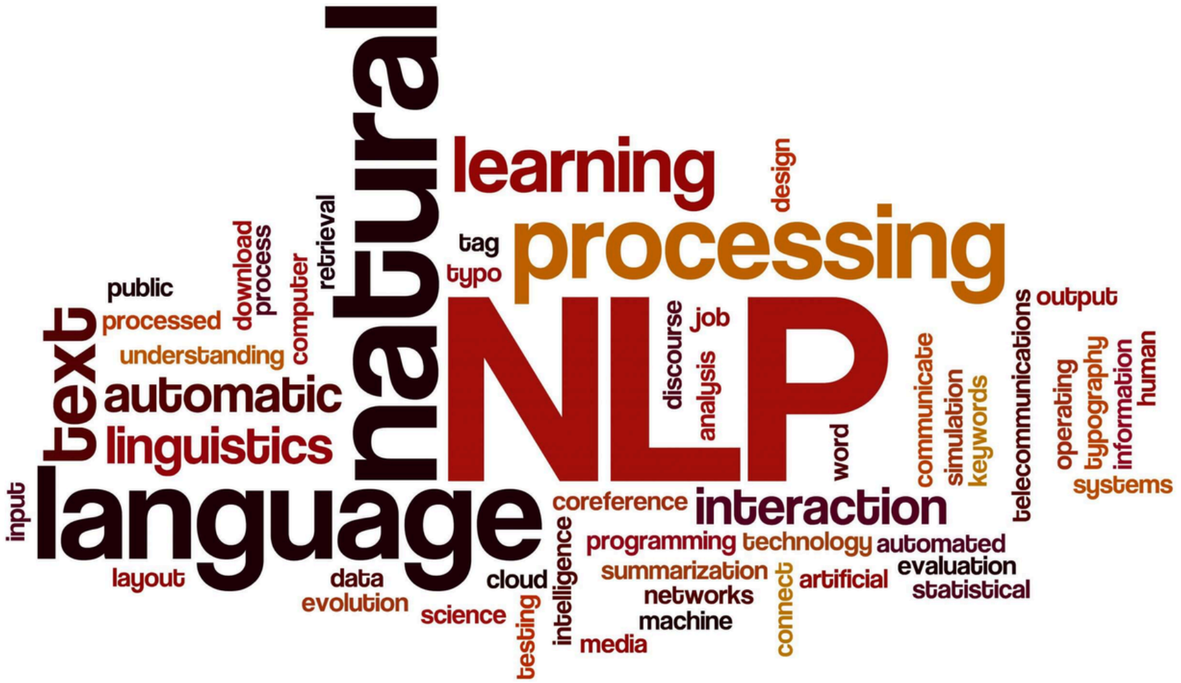
TOPIC MODELING WITH LDA & PROD-LDA

CORTINOVIS NICOLA, CORTOLEZZIS EDOARDO, LUCAS MARTA
UNIVERSITY OF TRIESTE

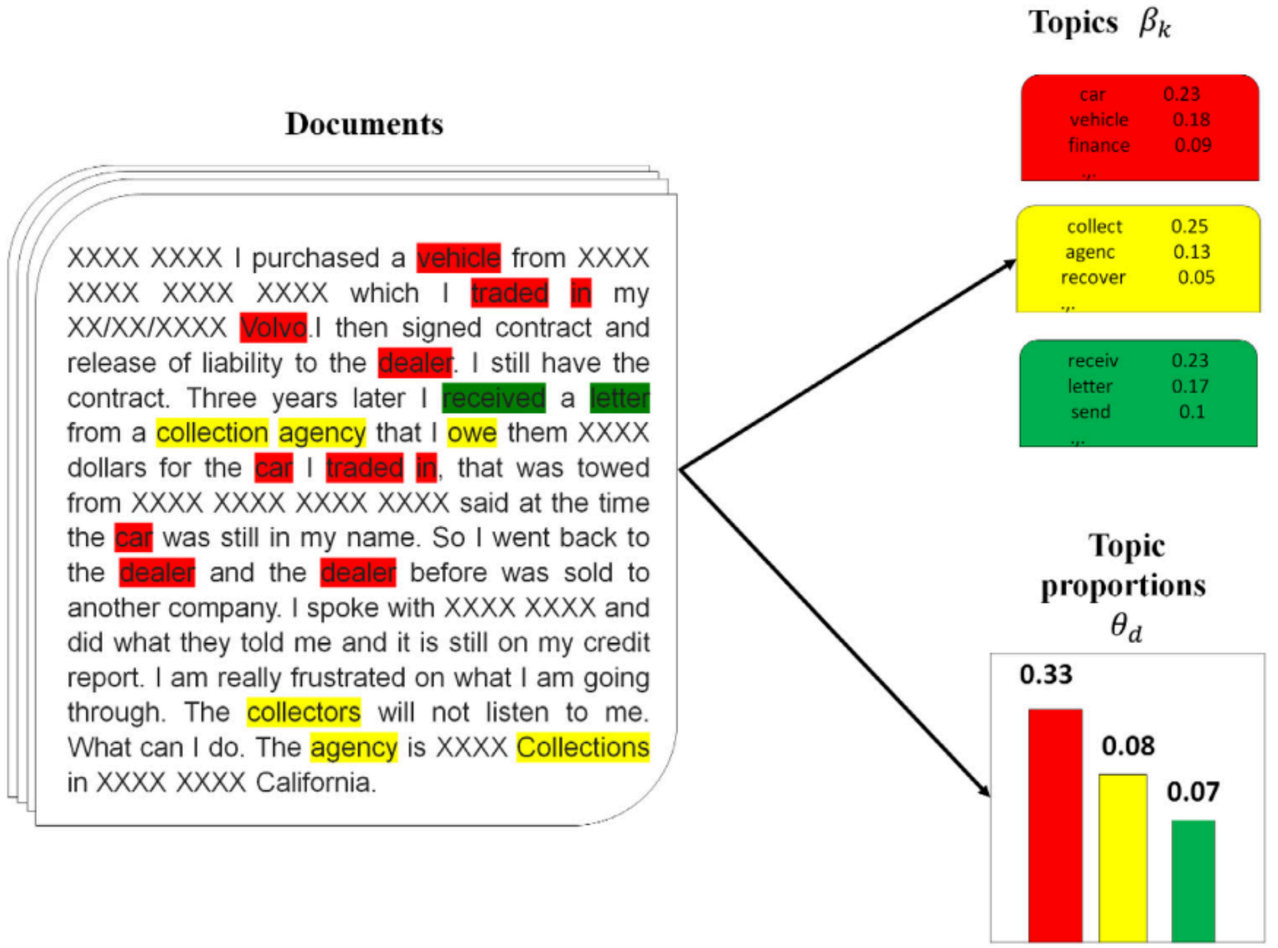
CONTENT

CONTENT



TOPIC MODELING

AI & HUMAN LANGUAGE



EXCHANGEABILITY

BOW

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

WORD

—

DOCUMENT

CORPUS

w_i $w = (w_1, w_2, \ldots, w_N)$ $D = \{w_1, w_2, \ldots, w_M\}$

CONTENT

DE FINETTI

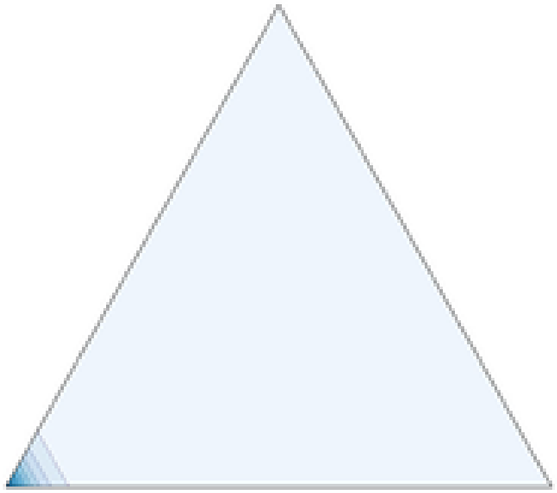
- EXCHANGEABLE RANDOM VARIABLES REPRESENTED AS A MIXTURE DISTRIBUTION;
- JOINT DISTRIBUTION OF AN INFINITELY EXCHANGEABLE RANDOM VARIABLES:
  A. RANDOM PARAMETER WERE DRAWN FROM SOME DISTRIBUTION
  B. RANDOM VARIABLES IID CONDITIONED ON THAT PARAMETER.

$$P(X_1, X_2, \ldots, X_n) = \int \prod_{i=1}^{n} P(X_i|\theta)P(\theta)d\theta$$

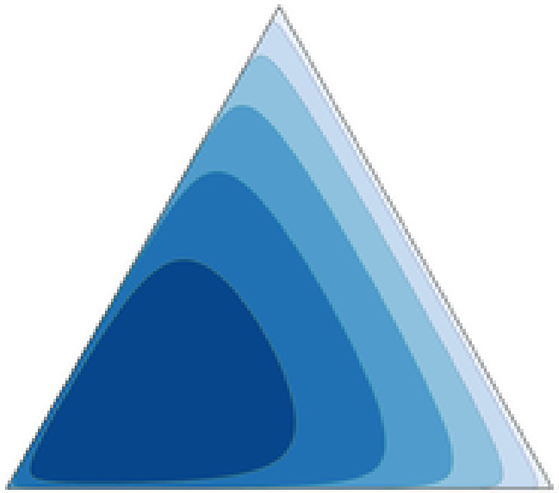
- IN LDA: WORDS GENERATED BY TOPICS , WHICH ARE INFINITELY EXCHANGEABLE WITHIN A DOCUMENT. ALSO, DOCUMENTS ARE EXCHANGABLE WITHIN DOCUMENTS.

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \,|\, \theta) p(w_n \,|\, z_n) \right) d\theta,$$

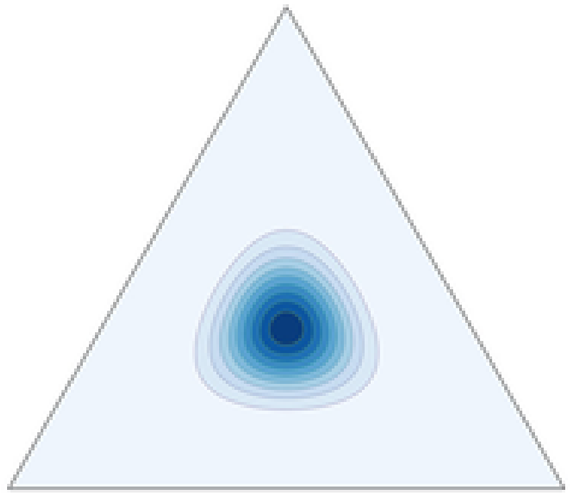
DIRICHLET DISTRIBUTION



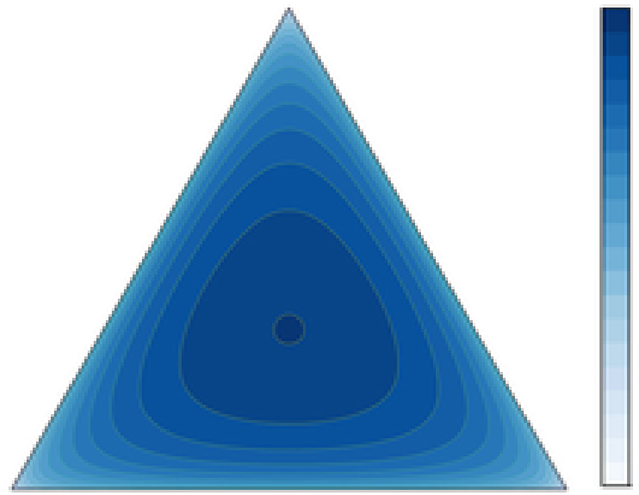
α = (20.0, 1.0, 1.0)



α = (1.5, 1.1, 1.1)



α = (10.0, 10.0, 10.0)



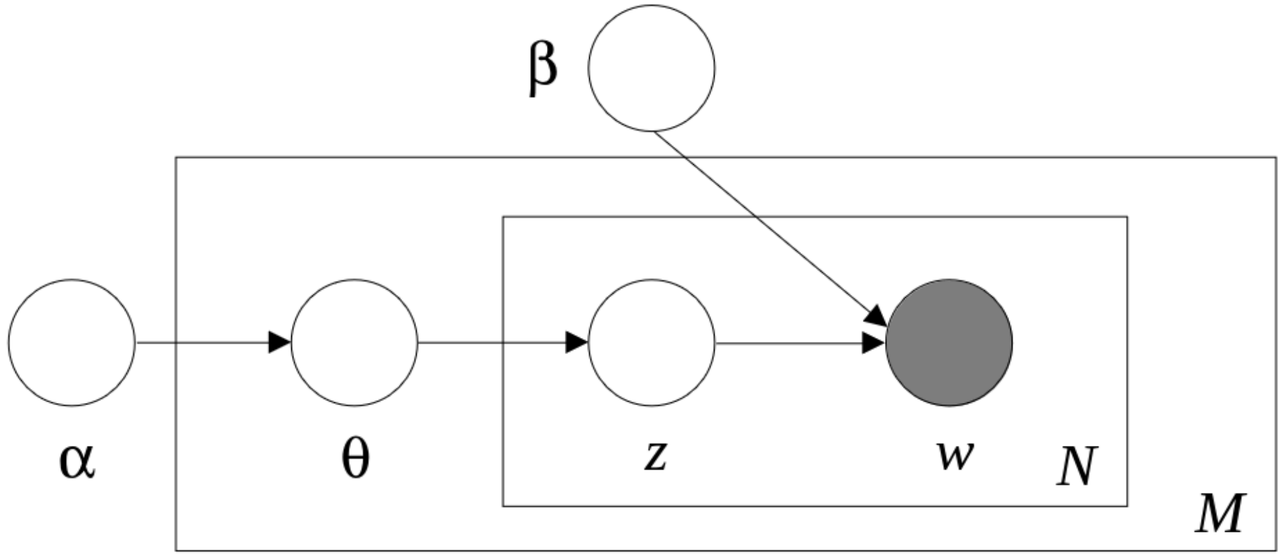
α = (1.1, 1.1, 1.1)

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}, \qquad \theta_i > 0, \quad \sum_{i=1}^{K}\theta_i = 1, \quad \forall i \in \{1,\ldots,K\}$$

CONTENT

HOW THE MODEL WORKS

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

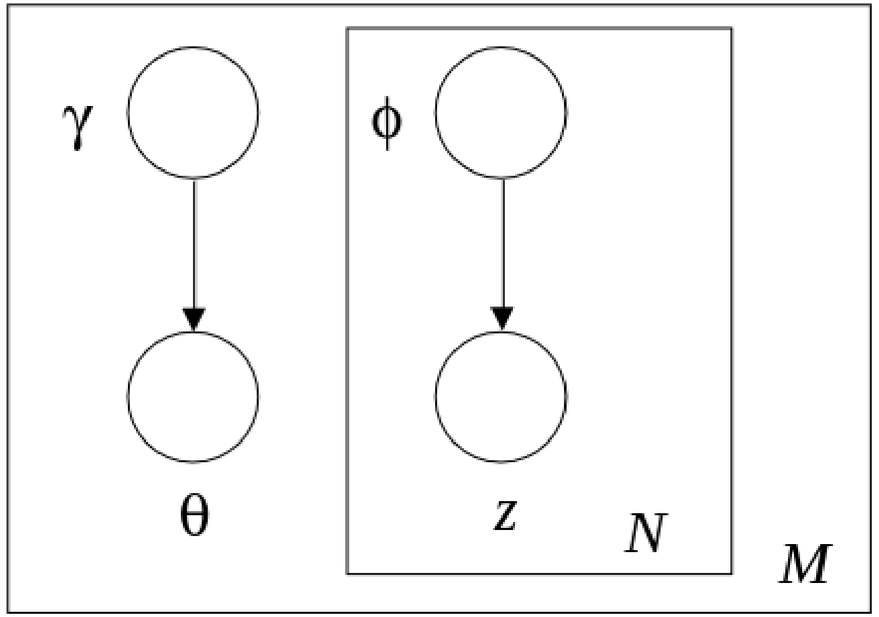


VARIATIONAL INFERENCE

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$



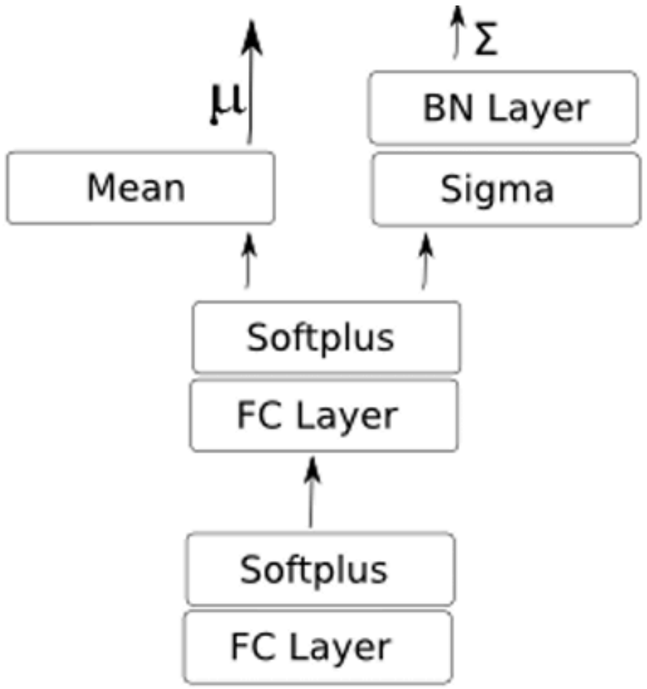
CONTENT

PROD-LDA

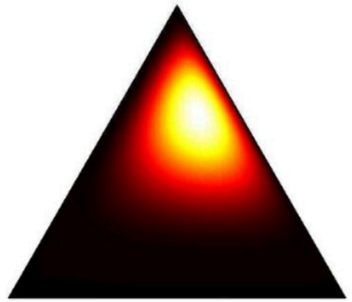
PRODUCT OF EXPERTS

$$p(w_n|\theta,\beta) \propto \prod_{k=1}^{K} p(w_n|z_n=k,\beta)^{\theta_k}$$

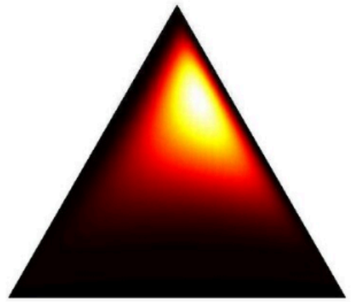
AVITM



$$\theta \sim \text{softmax}(\mathcal{N}(\mu,\Sigma))$$



Dirichlet

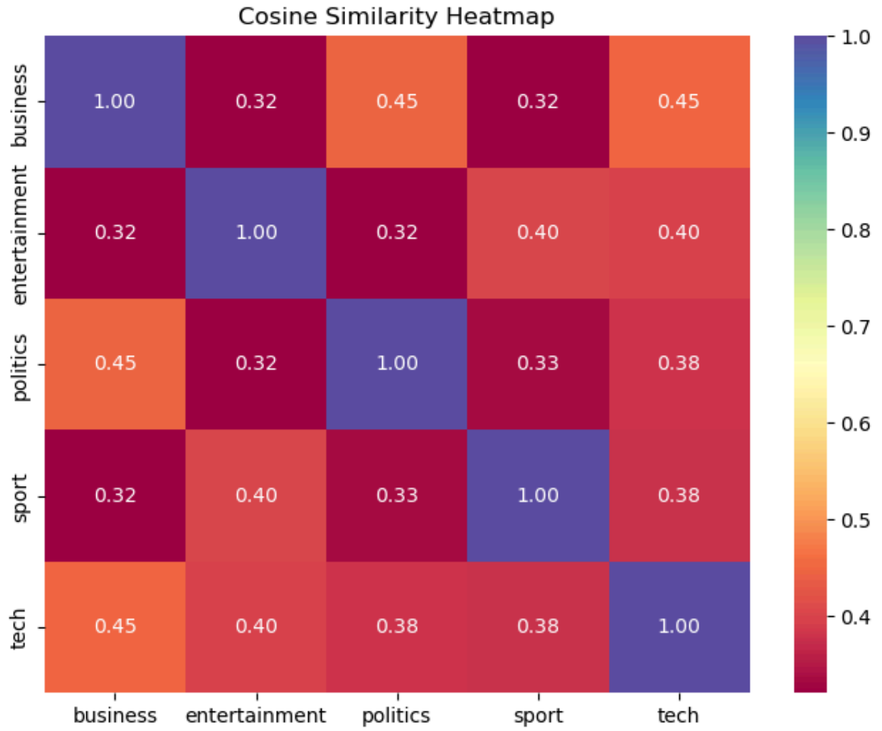


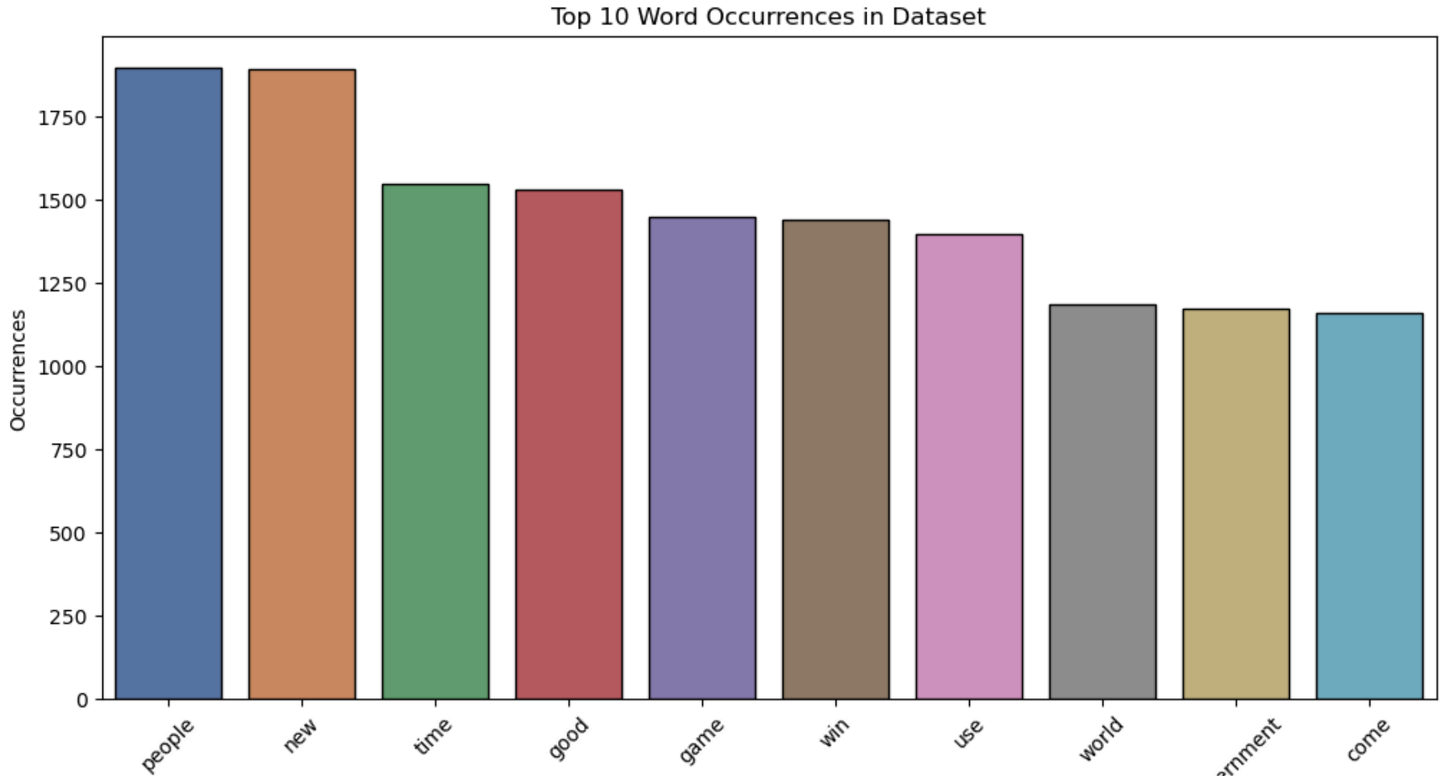
Logistic Normal

CONTENT

BBC NEWS DATASET OVERVIEW

TECH SPORT POLITICS ENTERTAINMENT BUSINESS





CONTENT

PIPELINE

LDA
PROD-LDA

→ RUNS →

BAYESIAN OPTIMIZATION
ON COHERENCE

↓

NUMBER OF TOPICS

NUMBER OF TOPICS
LEARNING RATE
MOMENTUM
NUMBER OF NEURONS

← RUNS ←

MODEL
SELECTION

METRICS

COHERENCE

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$

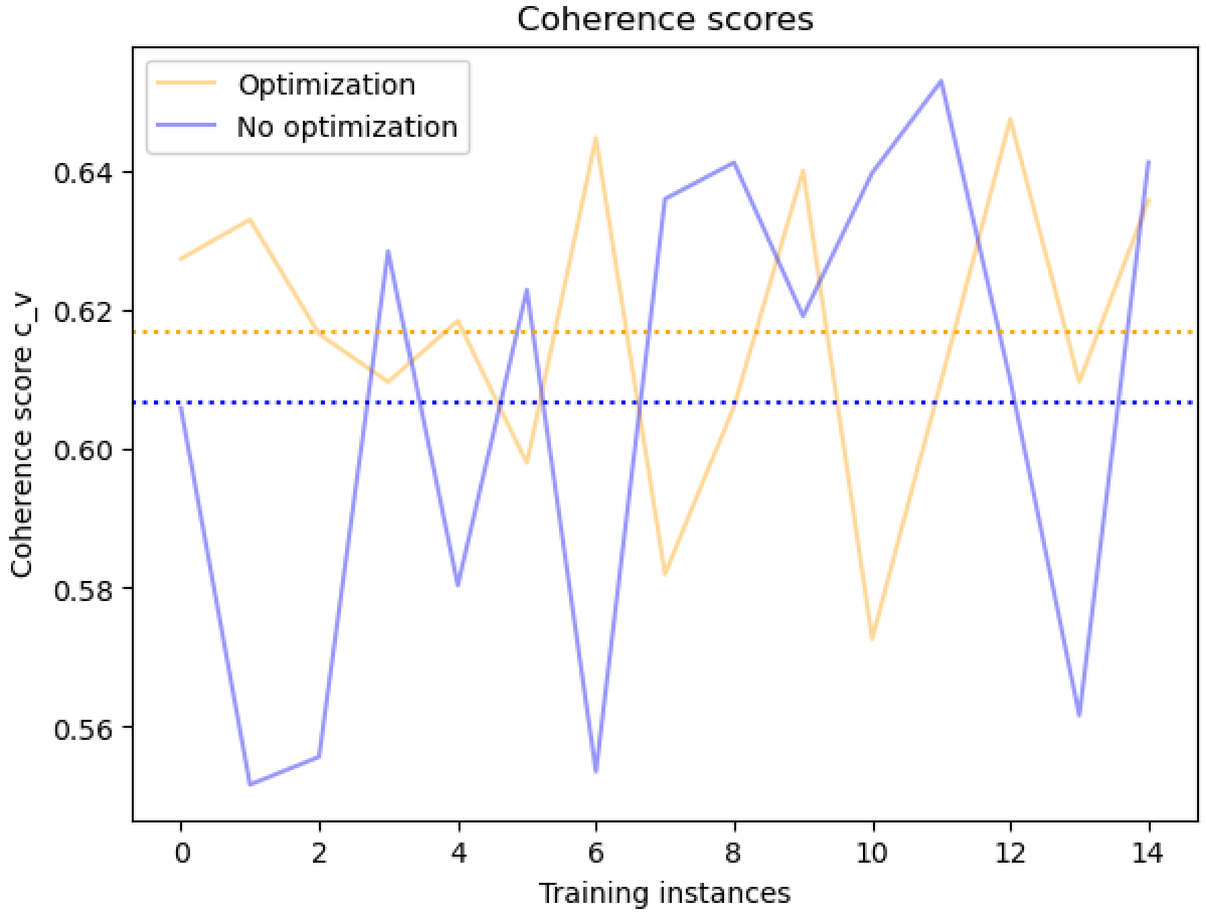
$$C(w_i) = \frac{1}{|P_i|} \sum_{w_j \in P_i} \mathrm{NPMI}(w_i, w_j)$$

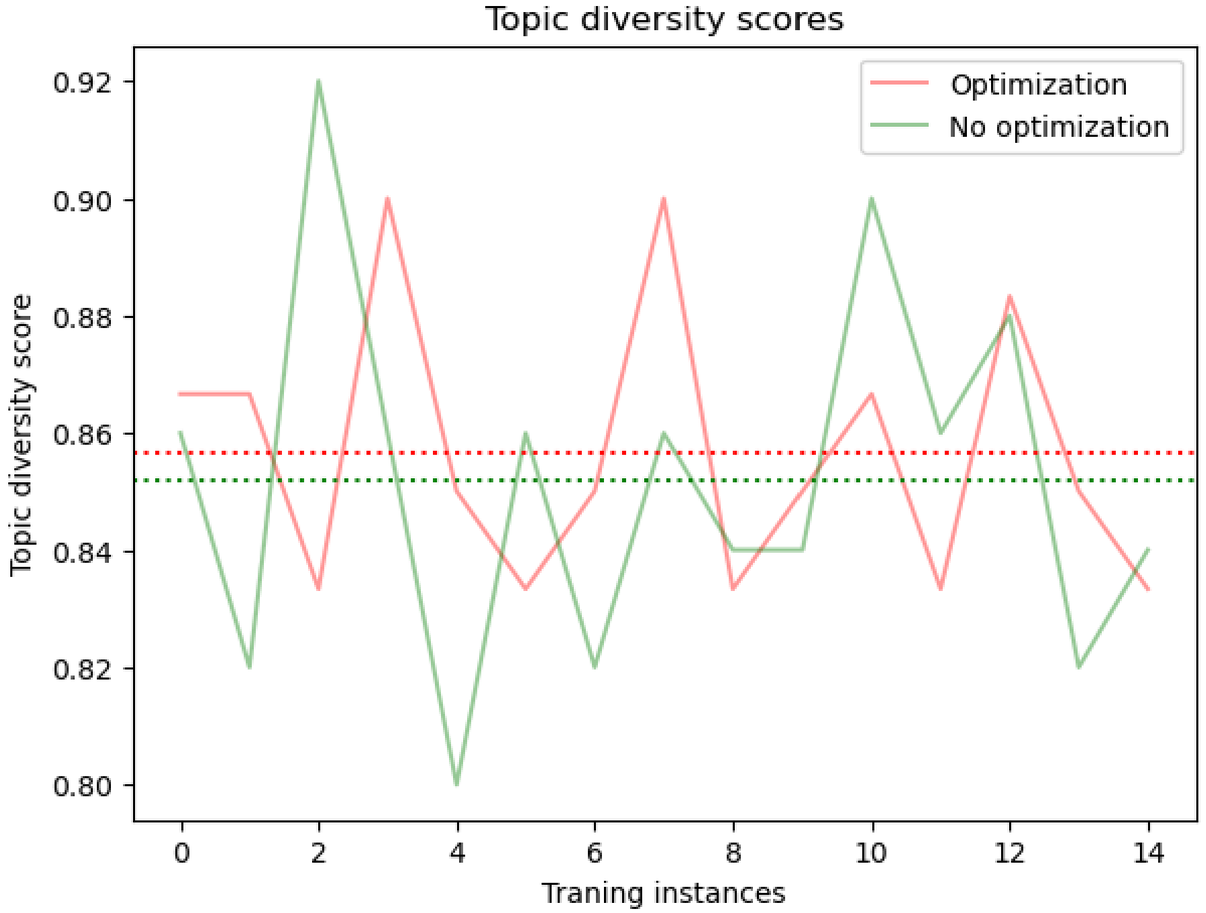
$$C_v(t) = \frac{1}{N} \sum_{i=2}^{N} C(w_i)$$

TOPIC DIVERSITY

$$\mathrm{TD} = \frac{|\bigcup_{t=1}^{T} W_t|}{T \times N} \times 100$$

LDA RESULTS





NUMBER OF TOPICS = 5

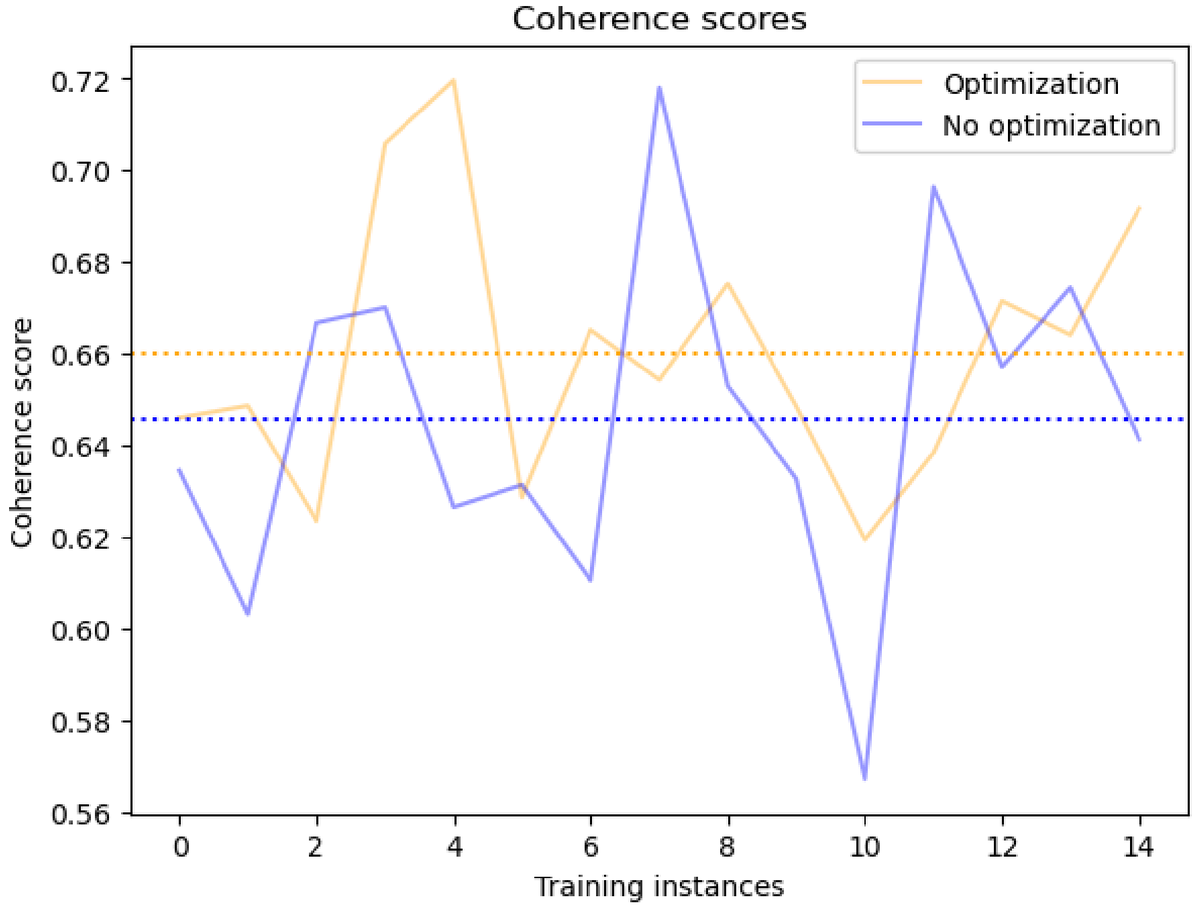
NON OPTIMIZED

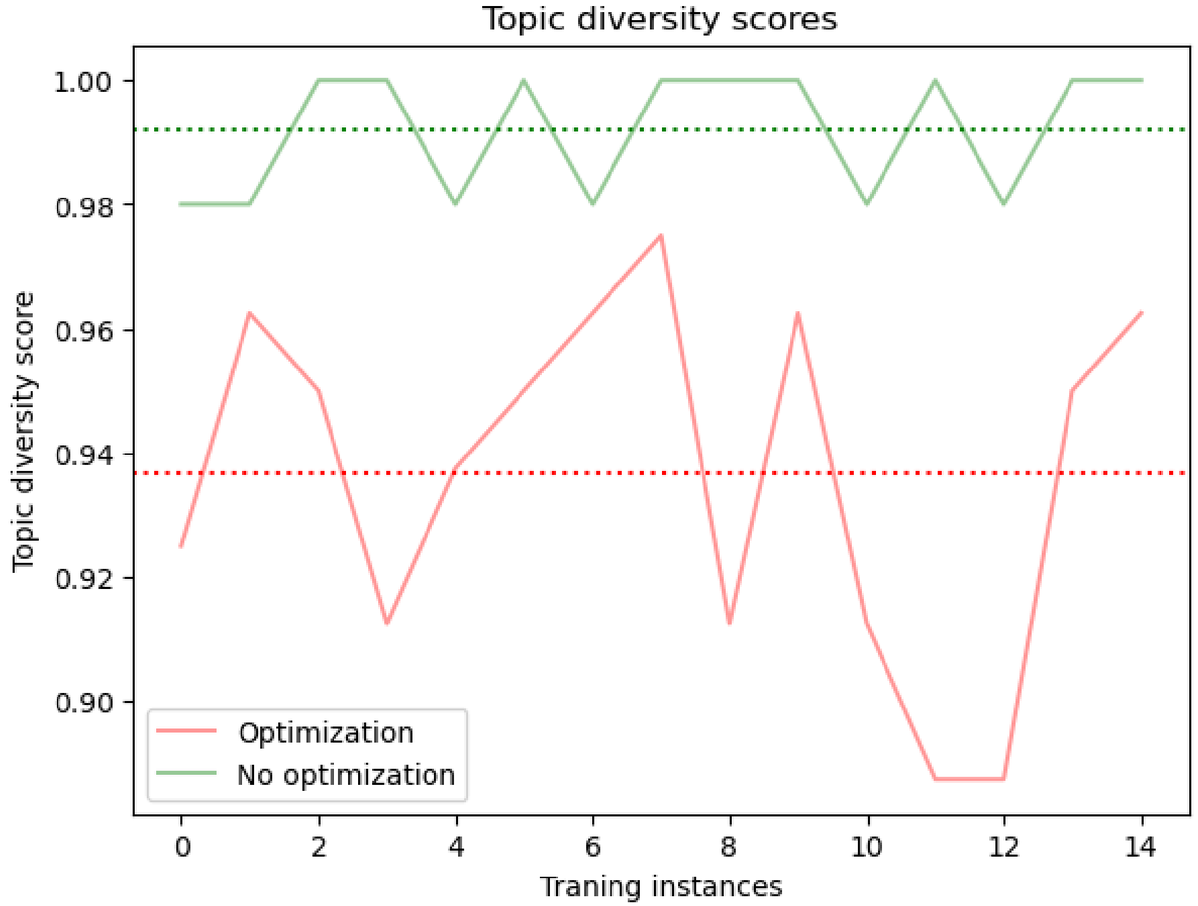
VS

OPTIMIZED

NUMBER OF TOPICS = 6

PROD LDA RESULTS





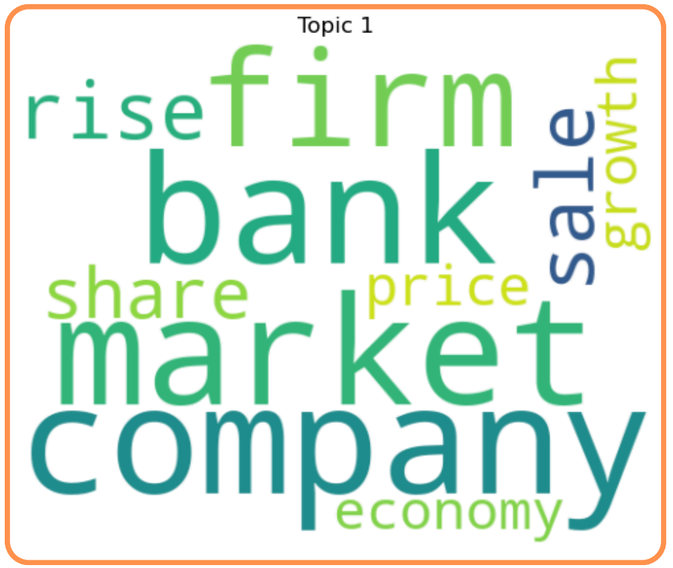
NUMBER OF TOPICS = 5
LR = 0.002
MOMENTUM = 0.99
NEURONS = 100

NON OPTIMIZED
VS
OPTIMIZED

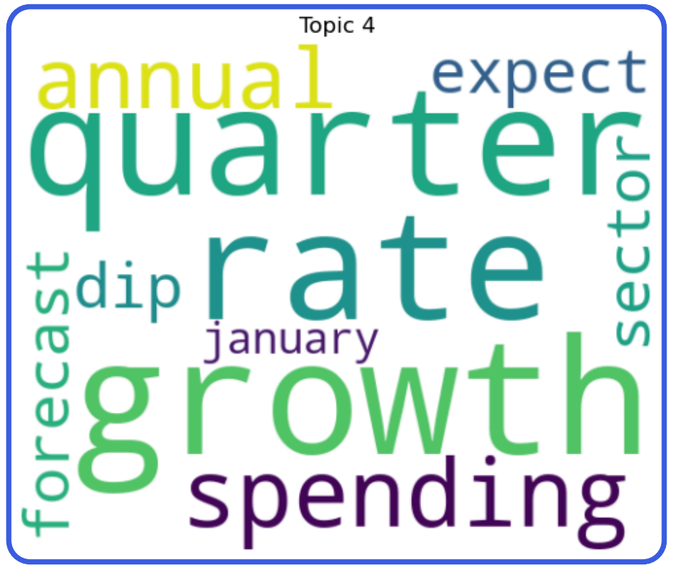
NUMBER OF TOPICS = 8
LR = 0.003
MOMENTUM = 0.99
NEURONS = 249

WORDCLOUDS COMPARISON

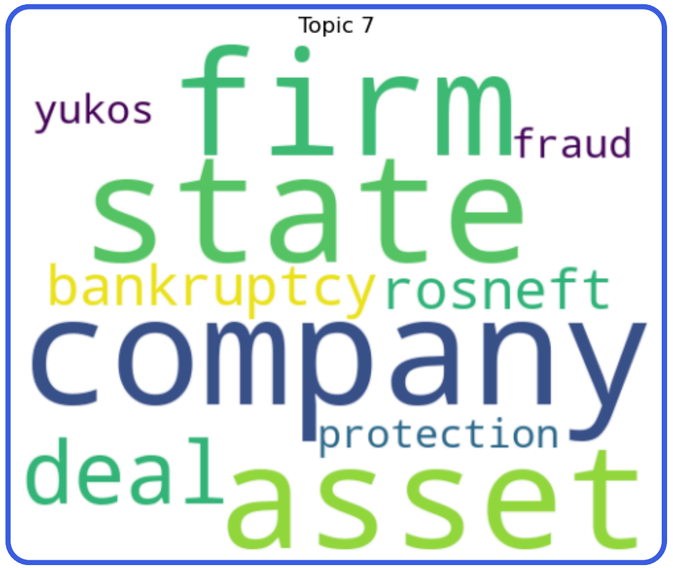
LDA



PROD-LDA



PROD-LDA



CONTENT

COMMENTS

PROD-LDA FINDS MORE SPECIFIC TOPICS AND THEY HAVE BETTER INTERPRETABILITY

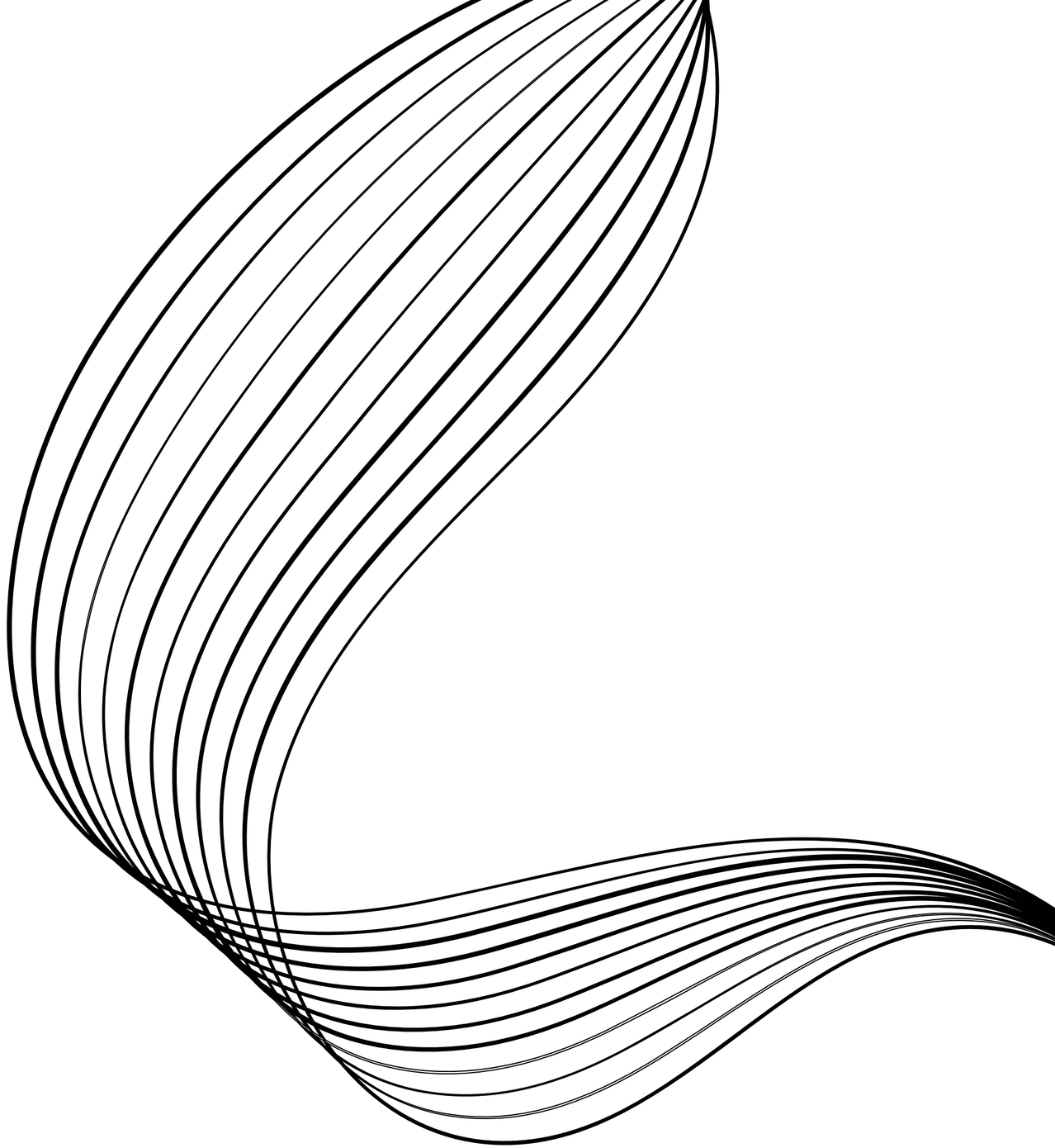
PROD-LDA IMPROVES COHERENCE, TOPIC DIVERSITY AND COMPUTATIONAL TIME

FUTURE WORKS

TO PRODUCE A MORE ACCURATE COMPARISON A MORE COMPLEX CORPUS SHOULD BE TESTED

THE BAYESIAN OPTIMIZATION ON THE HYPERPARAMETERS WAS LIMITED DUE TO THE HIGH COMPUTATIONAL DEMAND AND WAS NOT TOO IMPACTFUL IN FINDING BETTER MODEL

THANKS FOR YOUR ATTENTION!



REFERENCES

- Latent Dirichlet Allocation, M.Blei et alt. (2003);
- Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Jelodar et alt. (2018);
- Autoencoding Variational Inference for topic models, Srivastava and Sutton (2017);
- OCTIS: Comparing and Optimizing Topic Models is Simple!, Terragni et al. (2021)