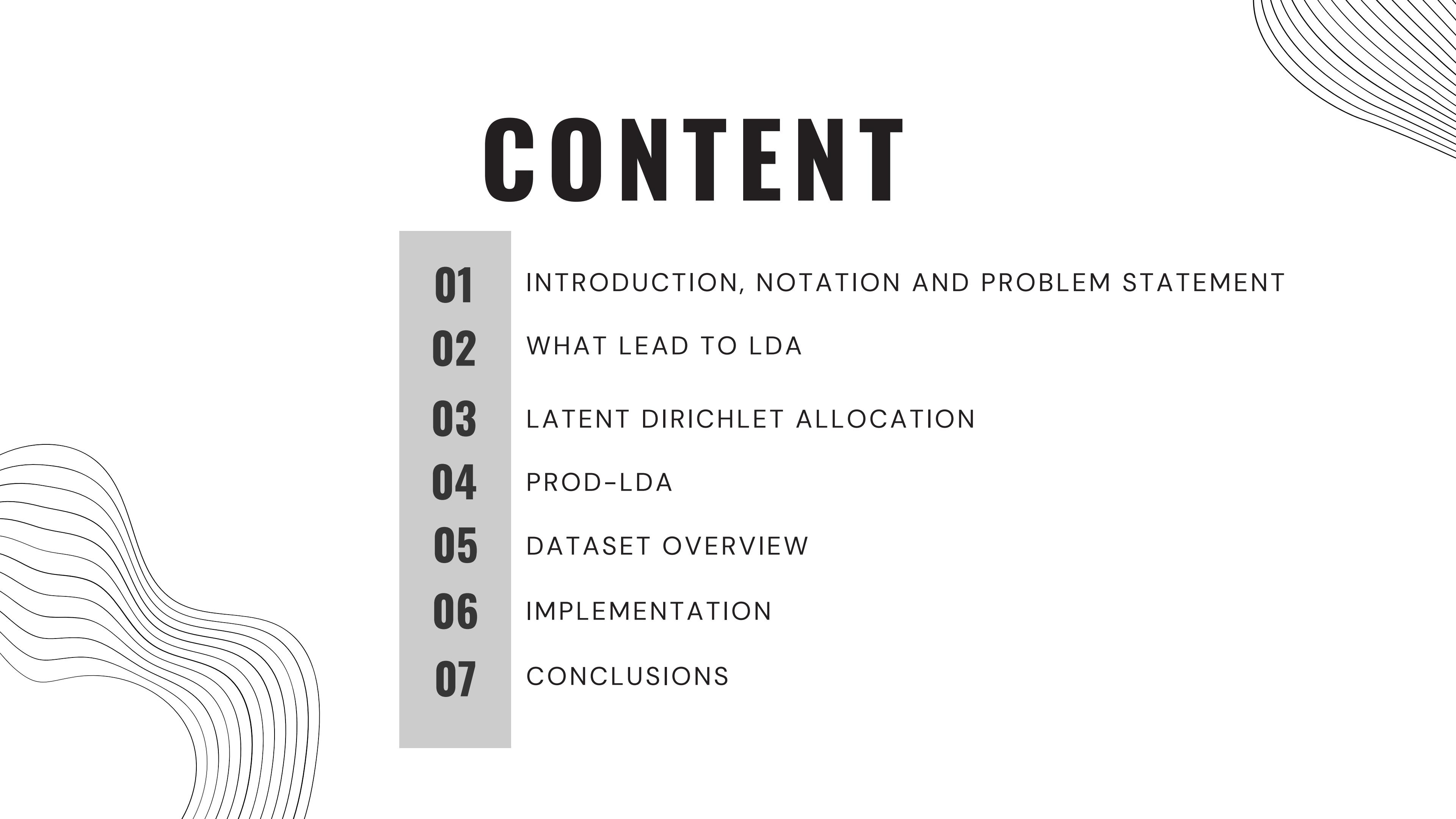




TOPIC MODELING WITH LDA & PROD-LDA

**CORTINOVIS NICOLA, CORTOLEZZIS EDOARDO, LUCAS MARTA
UNIVERSITY OF TRIESTE**

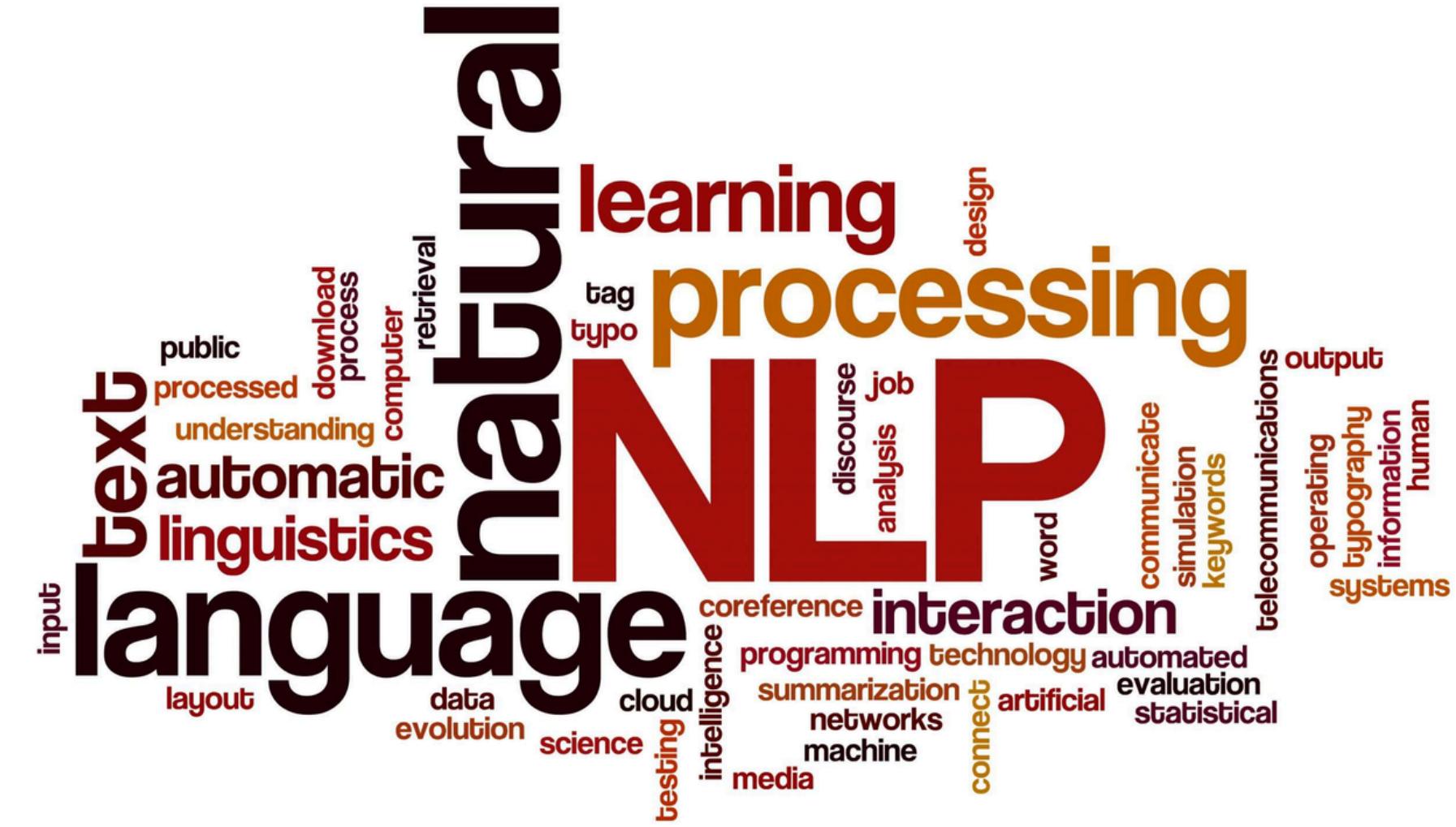
CONTENT

- 
- 01** INTRODUCTION, NOTATION AND PROBLEM STATEMENT
 - 02** WHAT LEAD TO LDA
 - 03** LATENT DIRICHLET ALLOCATION
 - 04** PROD-LDA
 - 05** DATASET OVERVIEW
 - 06** IMPLEMENTATION
 - 07** CONCLUSIONS

CONTENT

- 01**
- 02**
- 03**
- 04**
- 05**
- 06**
- 07**

INTRODUCTION, NOTATION AND PROBLEM STATEMENT



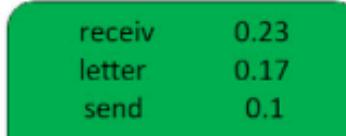
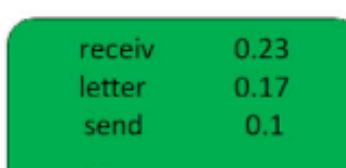
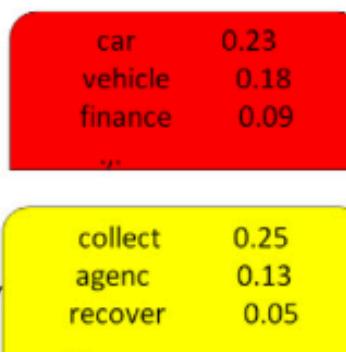
TOPIC MODELING

AI & HUMAN LANGUAGE

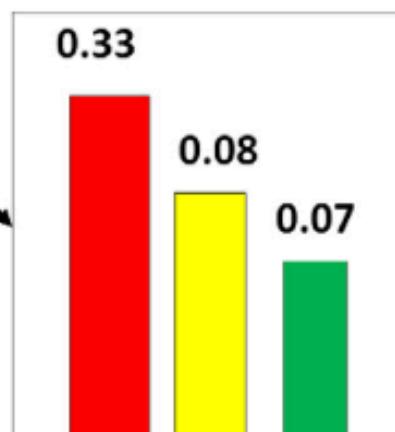
XXXX XXXX I purchased a vehicle from XXXX XXXX XXXX which I traded in my XX/XX/XXXX Volvo. I then signed contract and release of liability to the dealer. I still have the contract. Three years later I received a letter from a collection agency that I owe them XXXX dollars for the car I traded in, that was towed from XXXX XXXX XXXX XXXX said at the time the car was still in my name. So I went back to the dealer and the dealer before was sold to another company. I spoke with XXXX XXXX and did what they told me and it is still on my credit report. I am really frustrated on what I am going through. The collectors will not listen to me. What can I do. The agency is XXXX Collections in XXXX XXXX California.

Documents

Topics β_k



Topic proportions θ_d



EXCHANGEABILITY



WORD

—

w_i

DOCUMENT



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

BOW



CORPUS



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
...	...

CONTENT

01

02

03

04

05

06

07

WHAT LEAD TO LDA

DEFINETTI



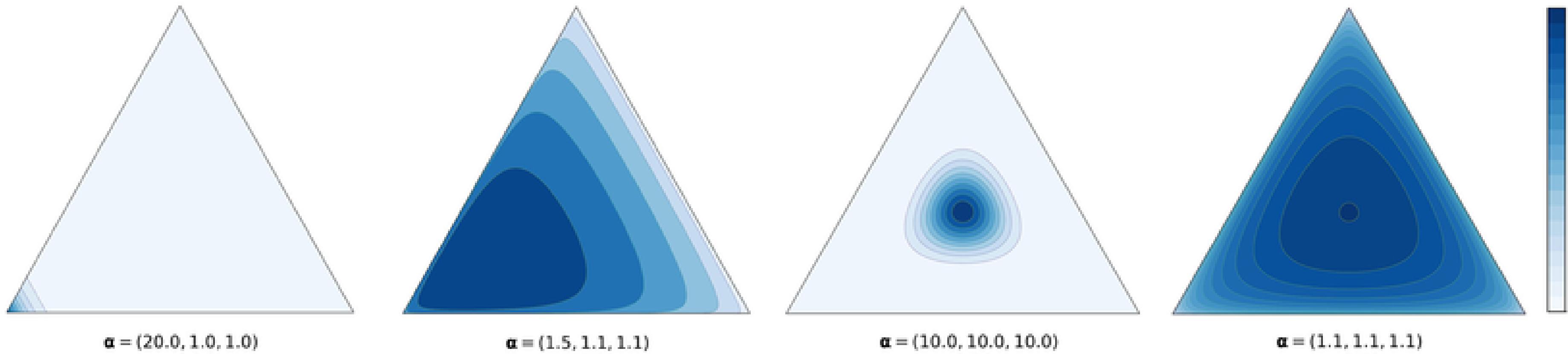
- EXCHANGEABLE RANDOM VARIABLES REPRESENTED AS A MIXTURE DISTRIBUTION;
- JOINT DISTRIBUTION OF AN INFINITELY EXCHANGEABLE RANDOM VARIABLES:
 - A. A RANDOM PARAMETER WERE DRAWN FROM SOME DISTRIBUTION
 - B. RANDOM VARIABLES IID CONDITIONED ON THAT PARAMETER.

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

- 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



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

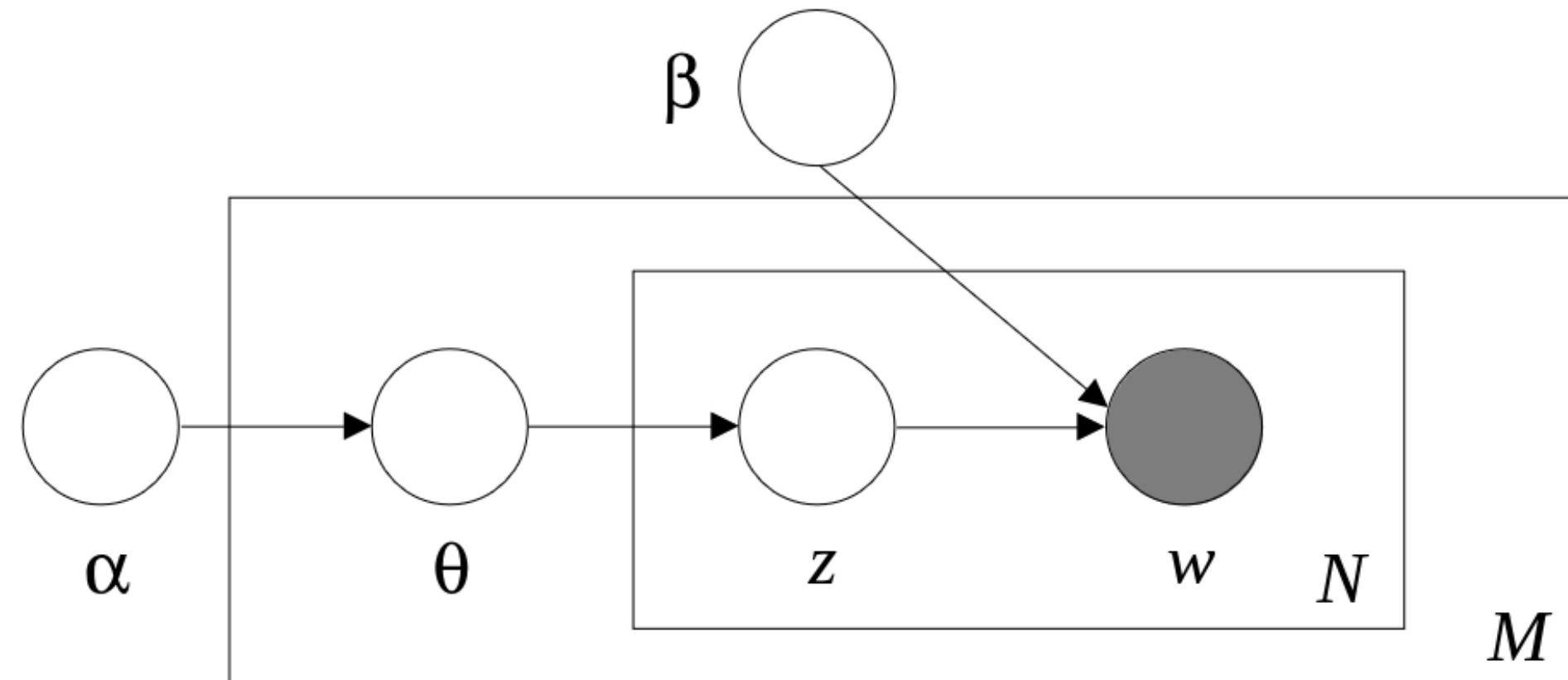
CONTENT

01
02
03
04
05
06
07

LATENT DIRICHLET ALLOCATION

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 $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n .

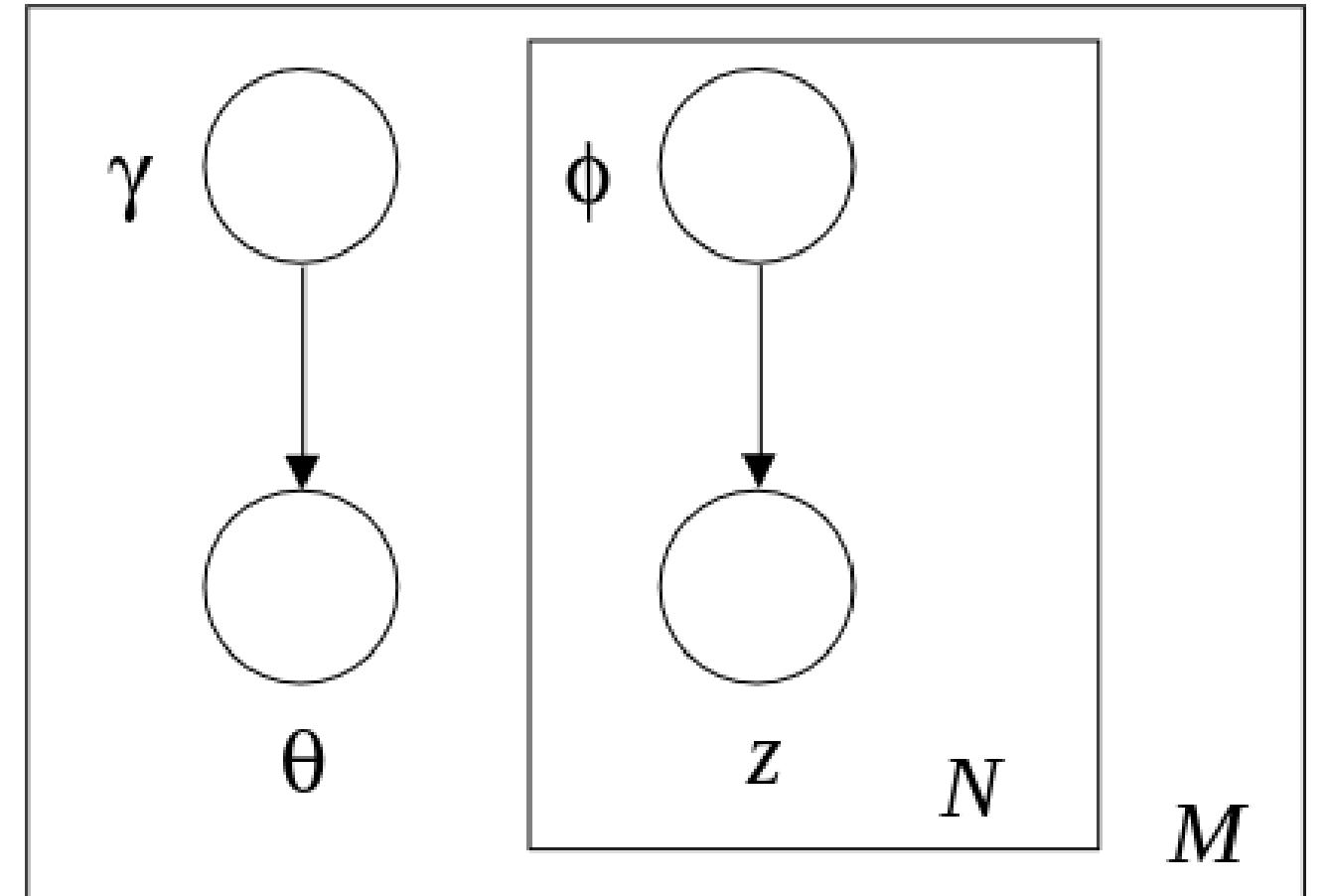


VARIATIONAL INFERENCE

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

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

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



CONTENT

01
02
03
04
05
06
07

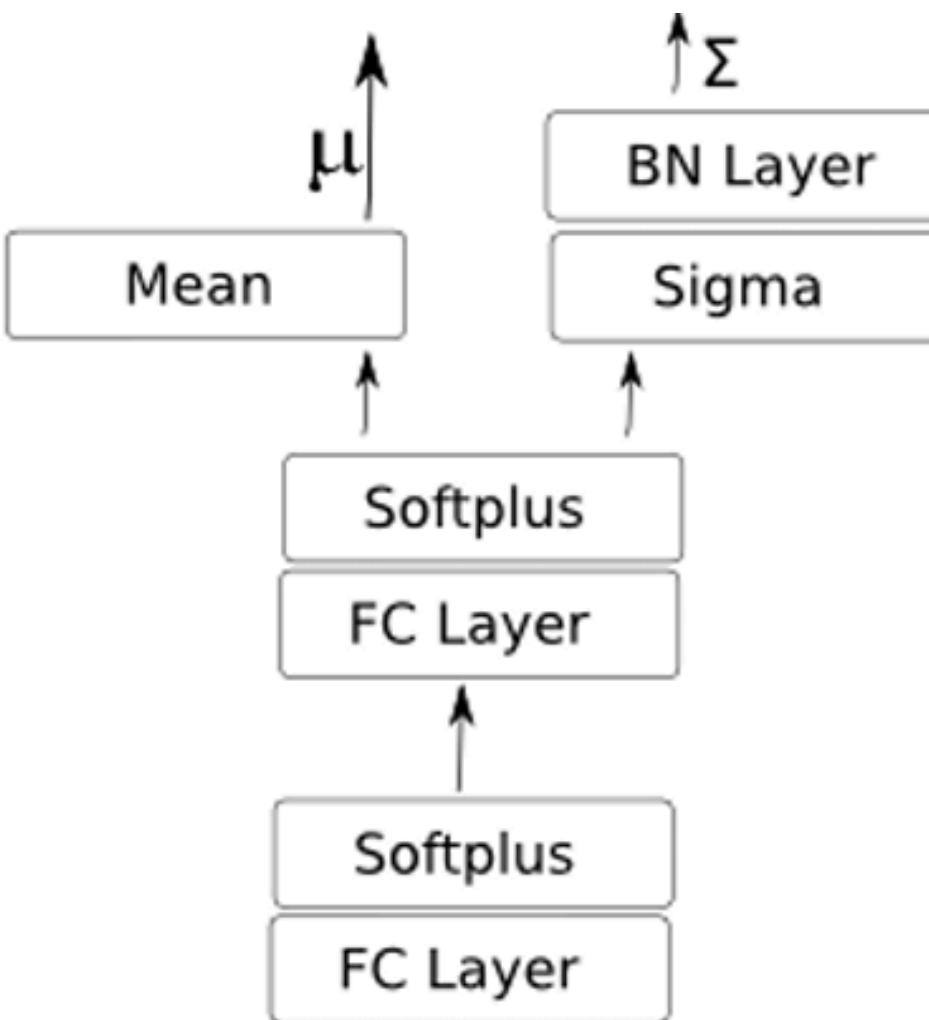
PROD-LDA

PROD-LDA

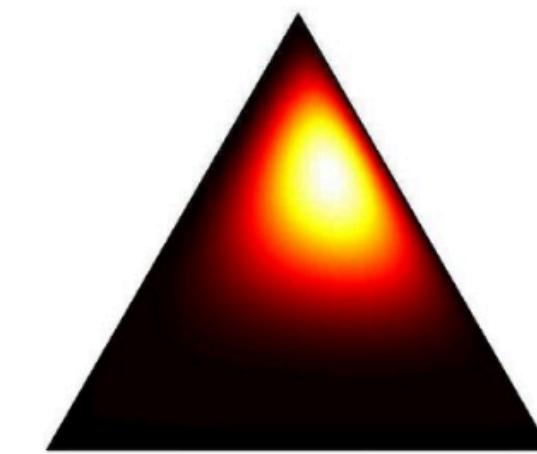
PRODUCT OF EXPERTS

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

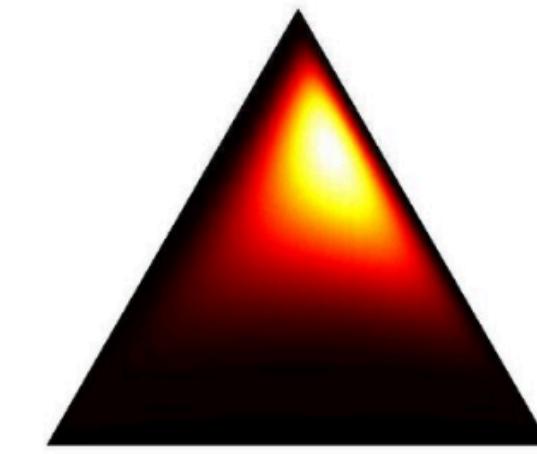
AVITM



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



Dirichlet



Logistic Normal

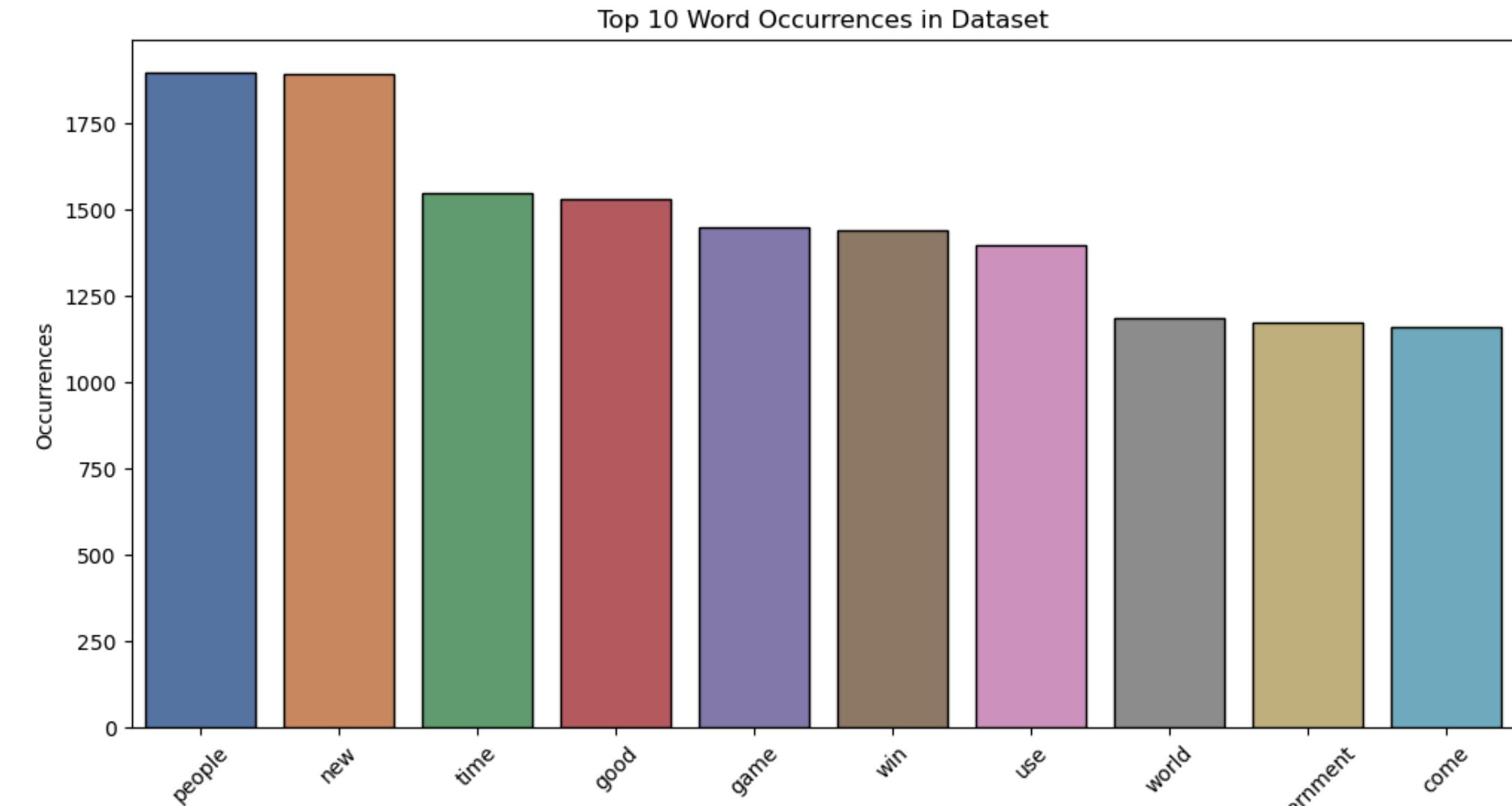
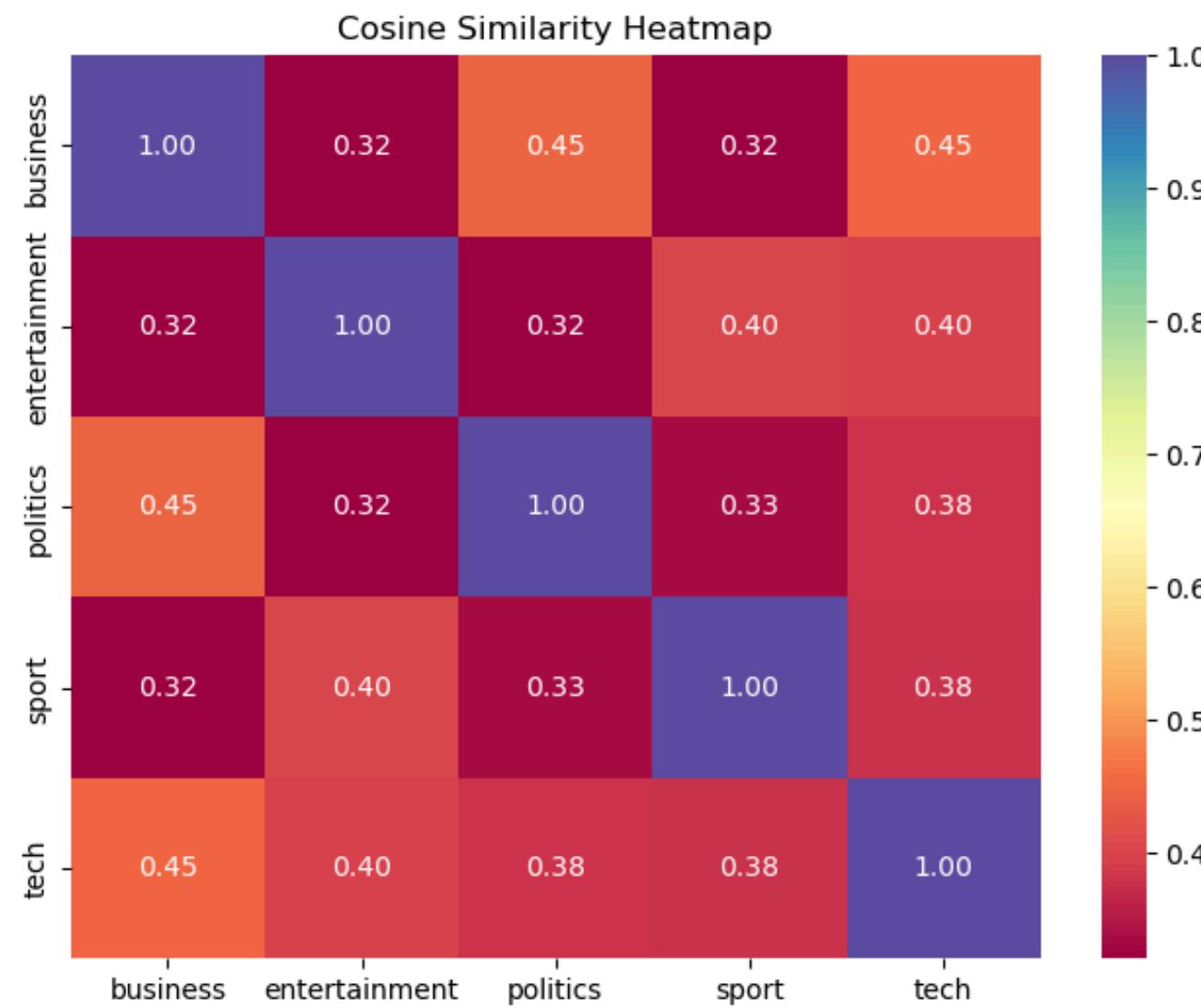
CONTENT

- 01
- 02
- 03
- 04
- 05
- 06
- 07

DATASET OVERVIEW

BBC NEWS DATASET OVERVIEW

TECH SPORT POLITICS ENTERTAINMENT BUSINESS

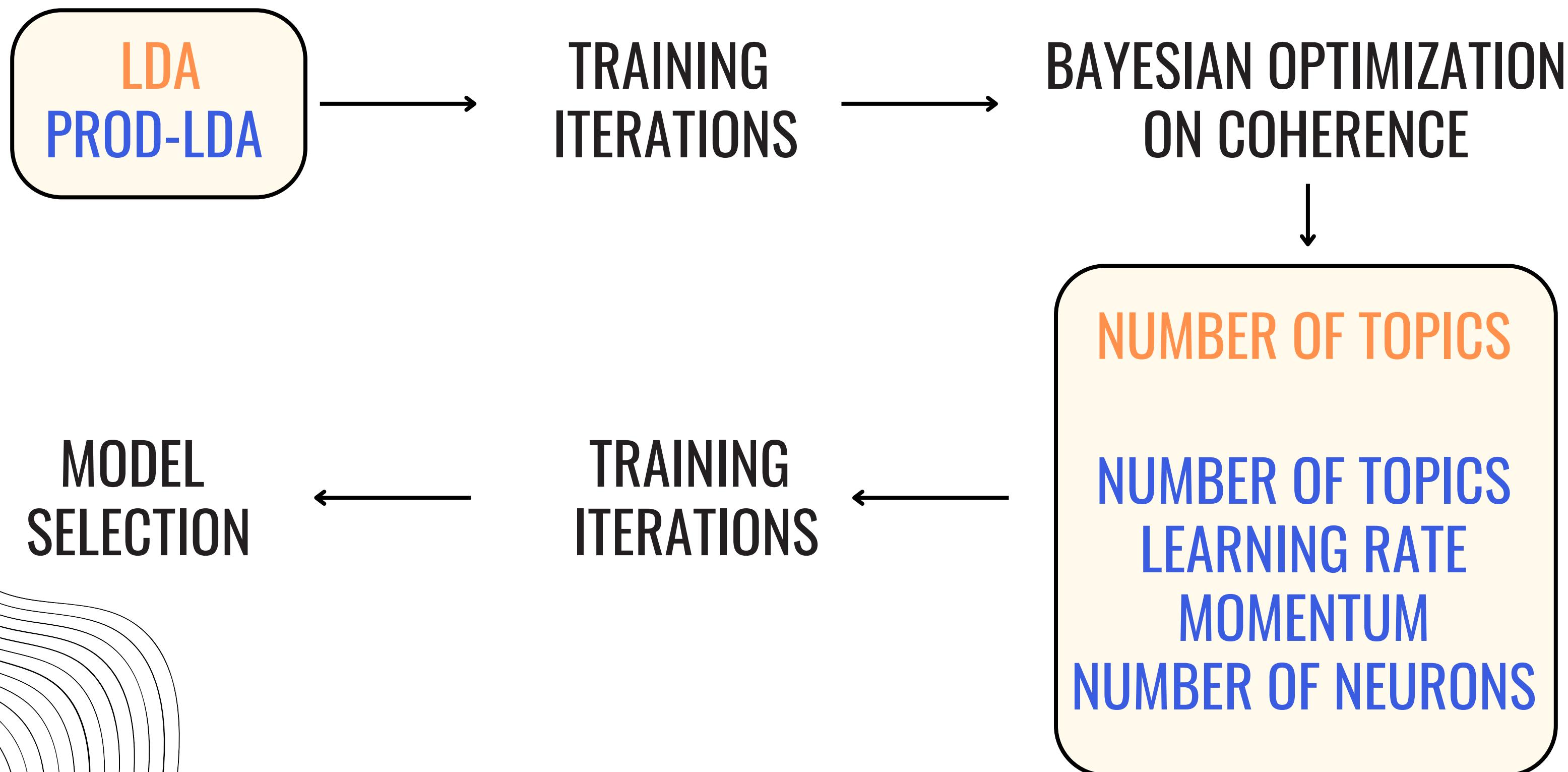


CONTENT

01
02
03
04
05
06
07

IMPLEMENTATION

PIPELINE



METRICS

COHERENCE

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

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

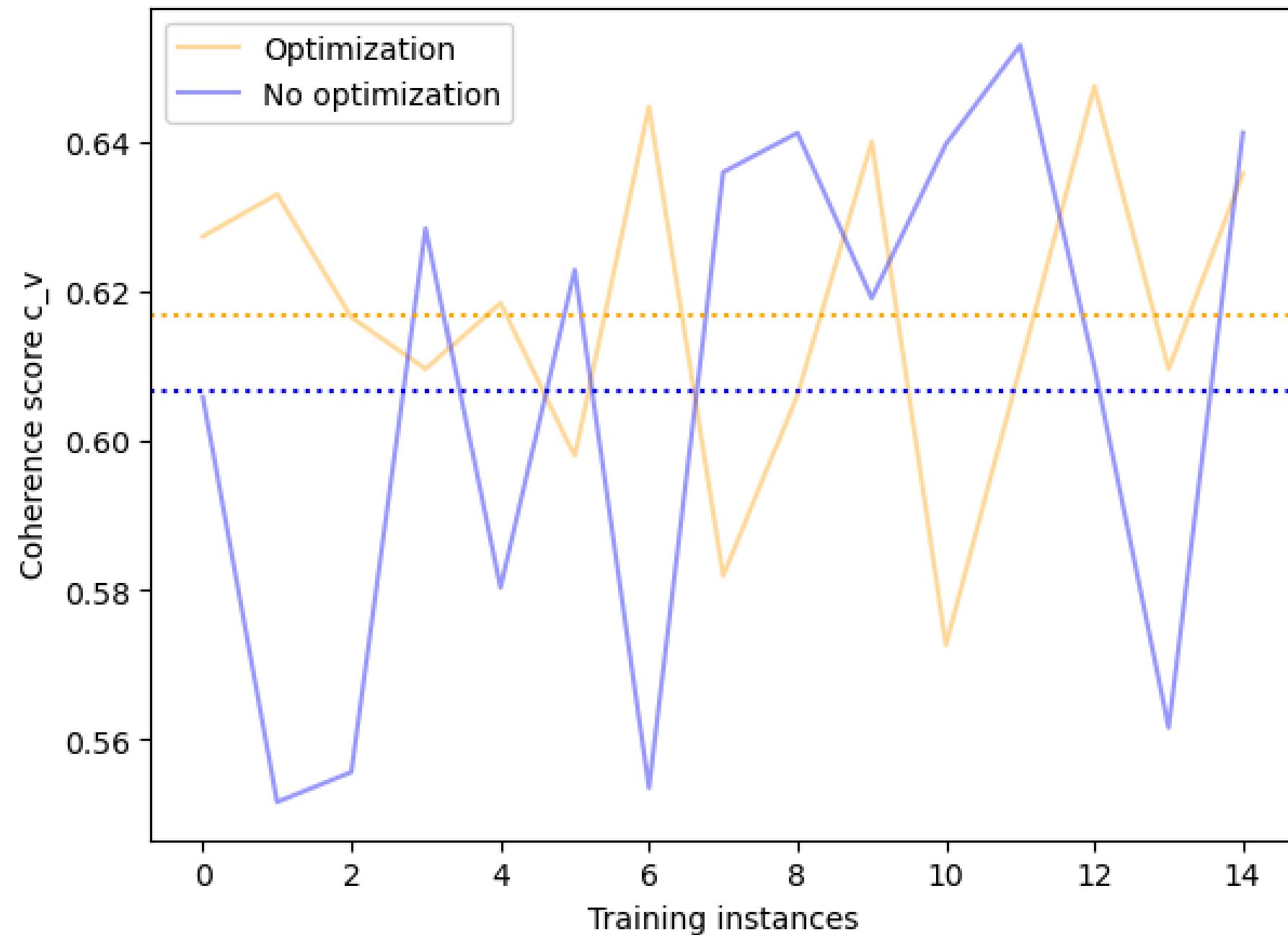
$$C_v(t) = \frac{1}{N} \sum_{i=2}^N C(w_i)$$

TOPIC DIVERSITY

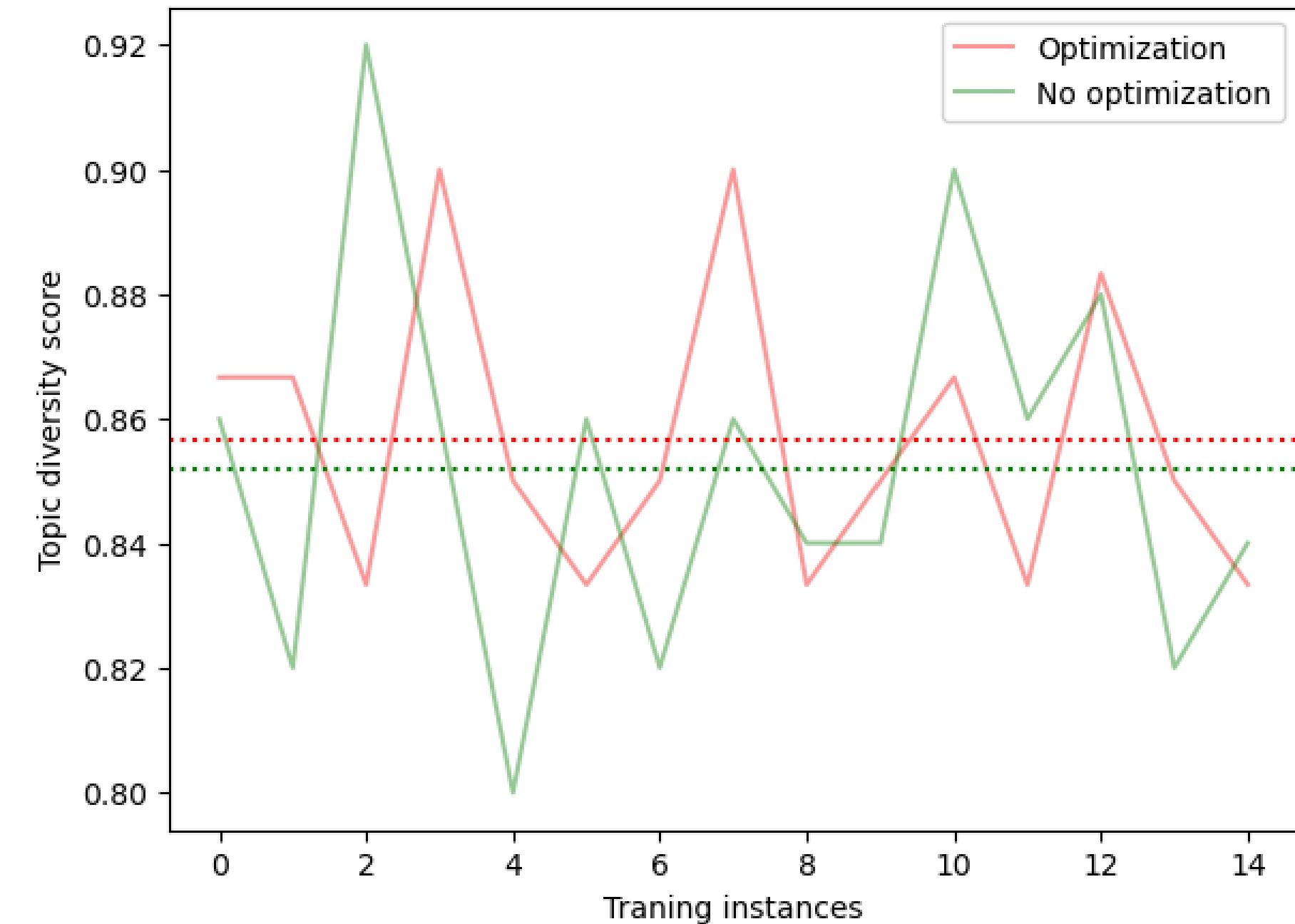
$$\text{TD} = \frac{|\bigcup_{t=1}^T W_t|}{T \times N} \times 100$$

LDA RESULTS

Coherence scores



Topic diversity scores

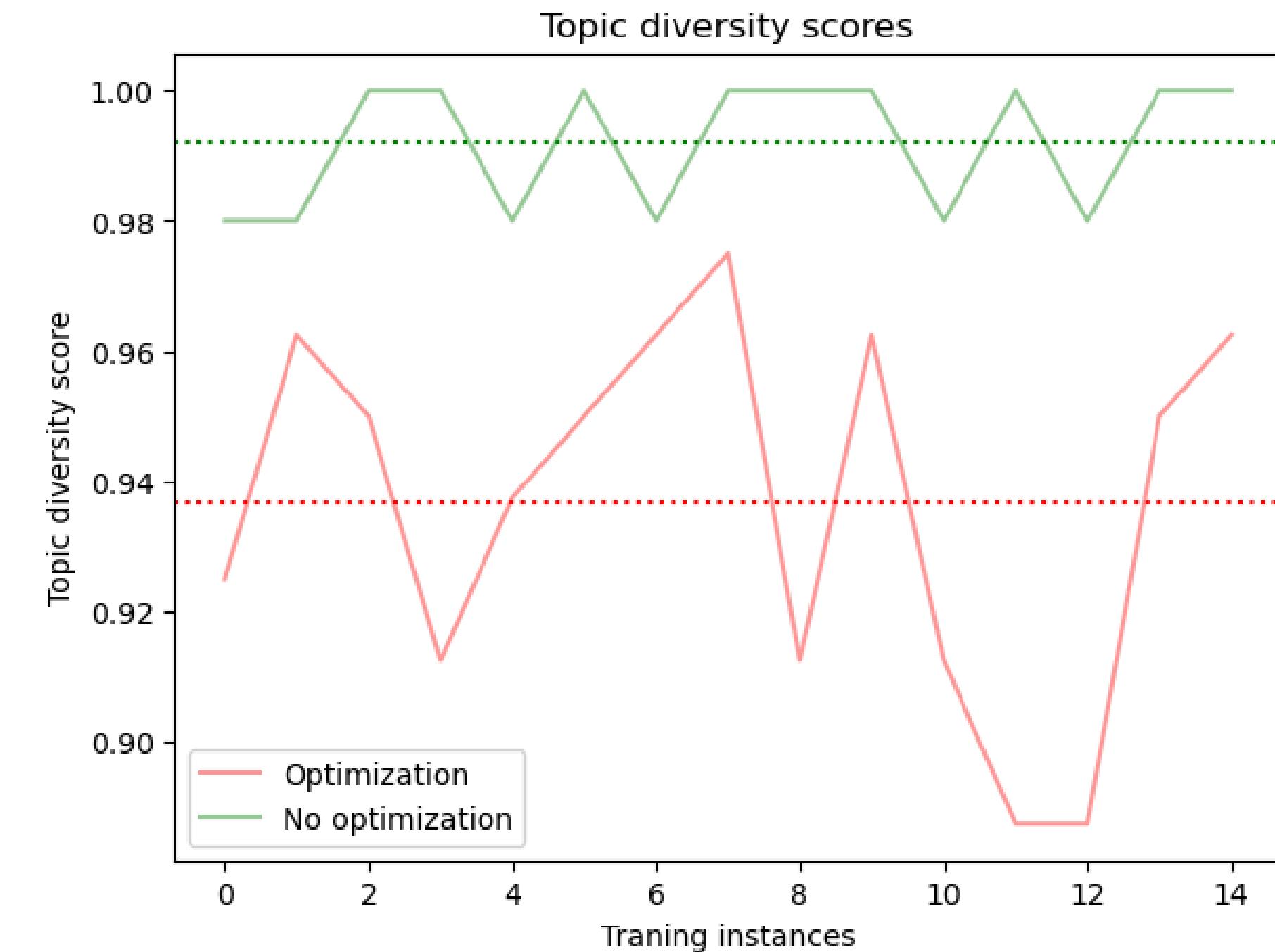
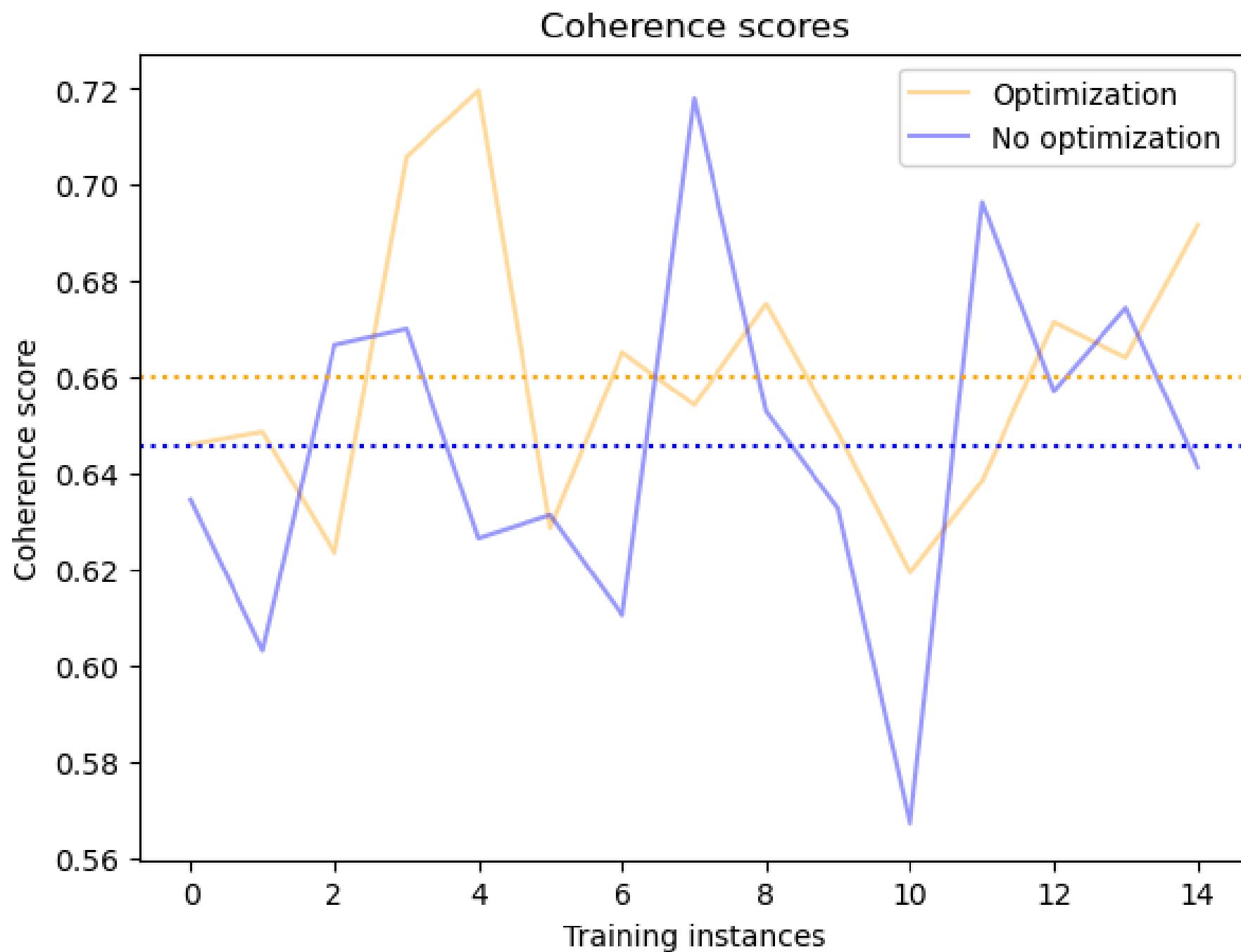


NUMBER OF TOPICS = 5

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

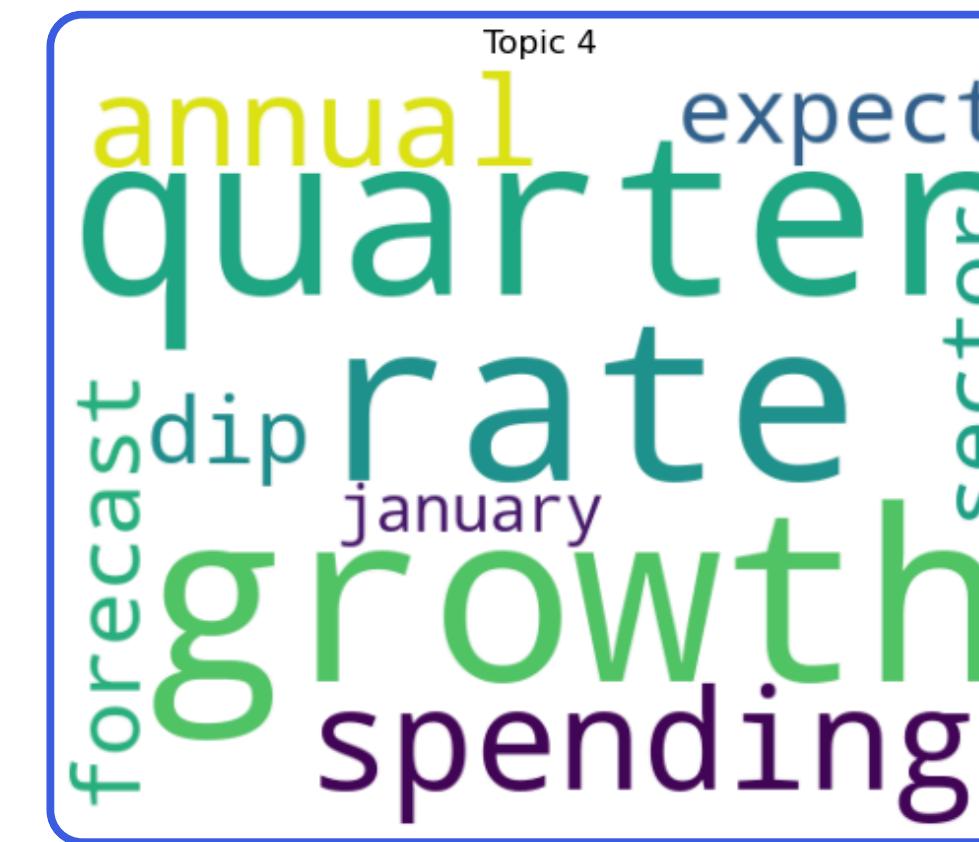
NUMBER OF TOPICS = 8
LR = 0.003
MOMENTUM = 0.99
NEURONS = 249

WORDCLOUDS COMPARISON

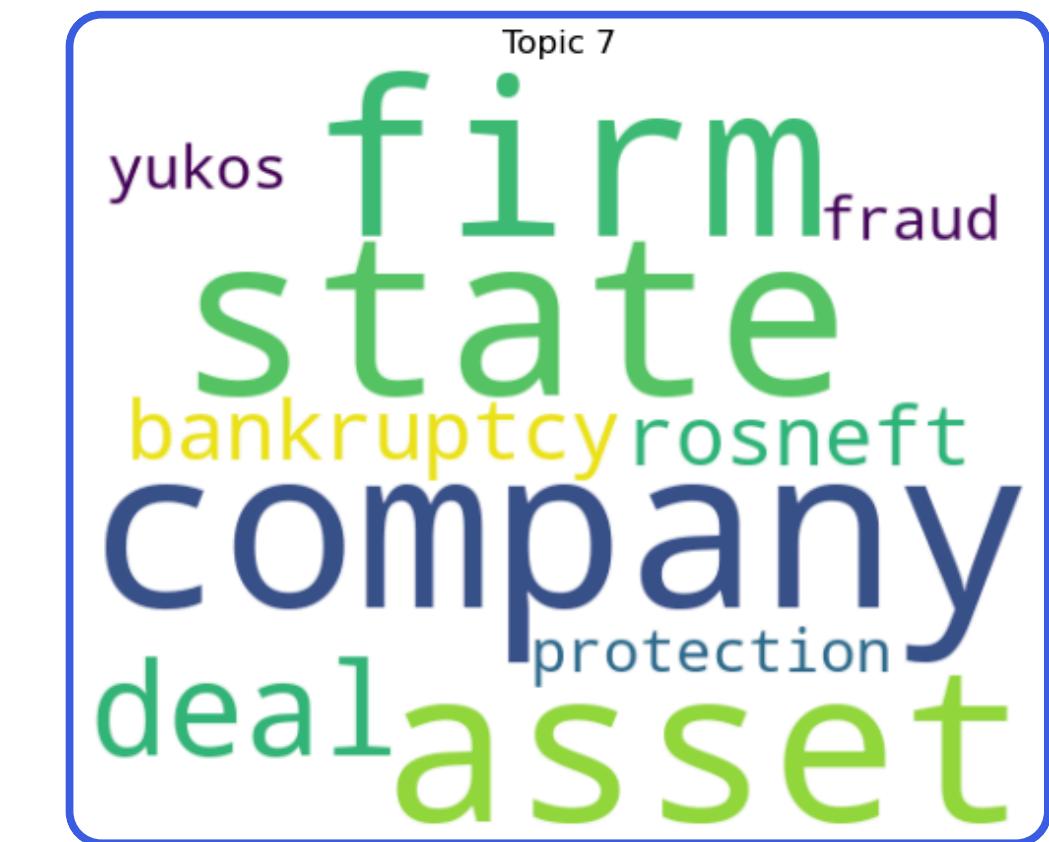
LDA



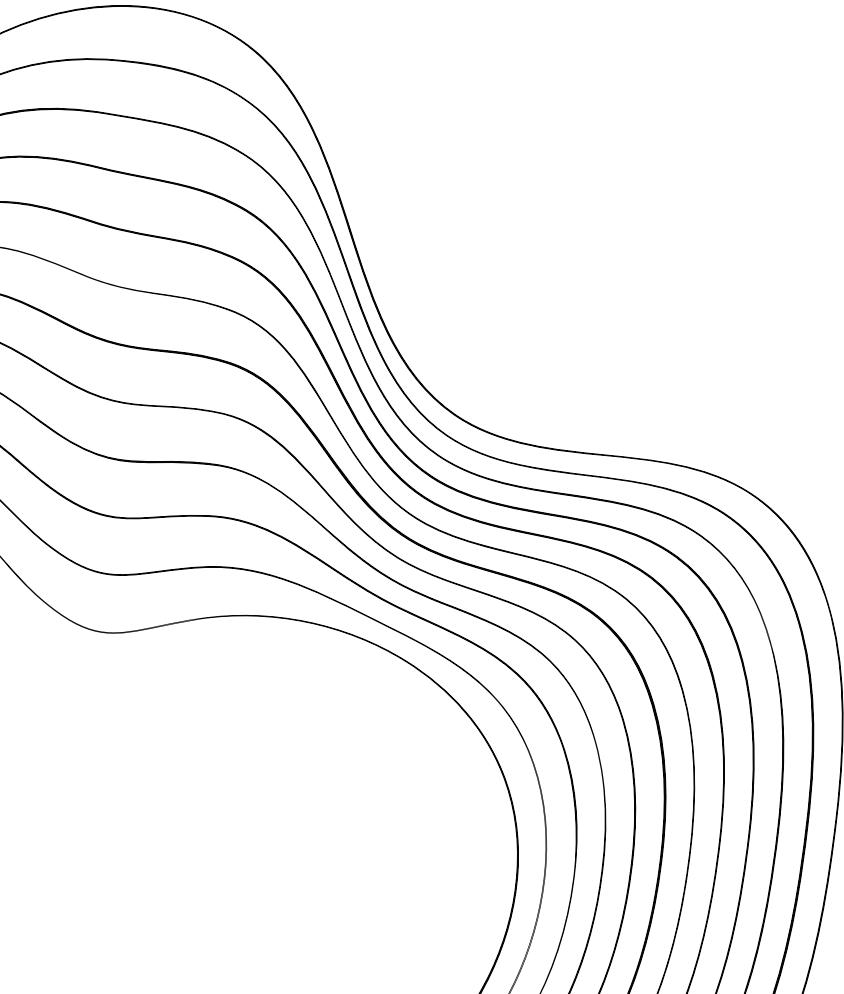
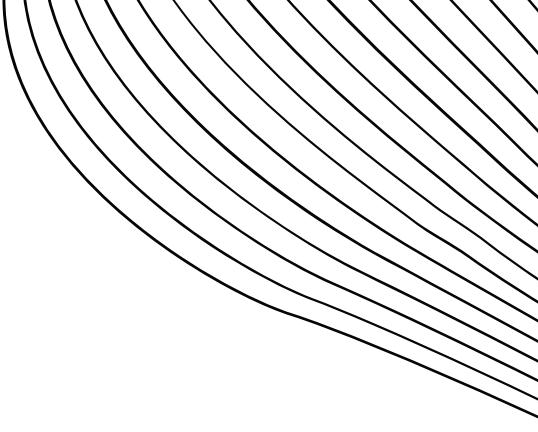
PROD-LDA



PROD-LDA



CONTENT

- 
- 
- 01
 - 02
 - 03
 - 04
 - 05
 - 06
 - 07

CONCLUSIONS

COMMENTS

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

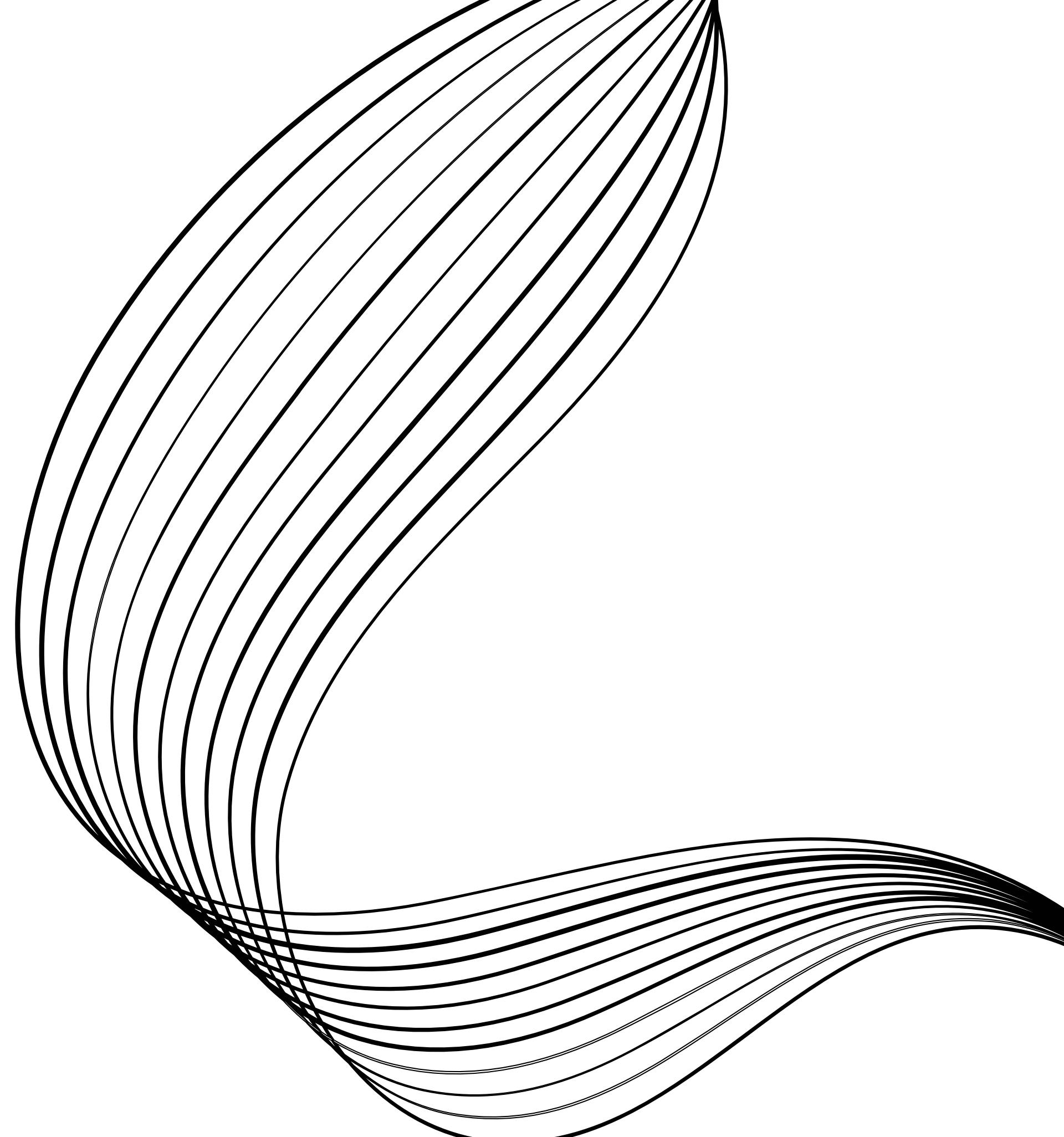
PROD-LDA IMPROVES COHERENCE,
TOPIC DIVERSITY AND COMPUTATIONAL
TIME

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)