Modeling research topics in movement ecology

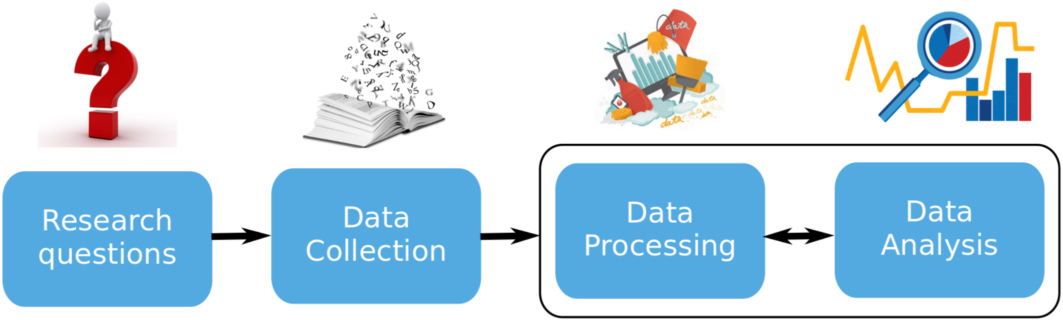
Rocío Joo
ME Boone, S Picardi, VS Romero-Romero, TA Clay, SC Patrick, M Basille

IBC - 2020

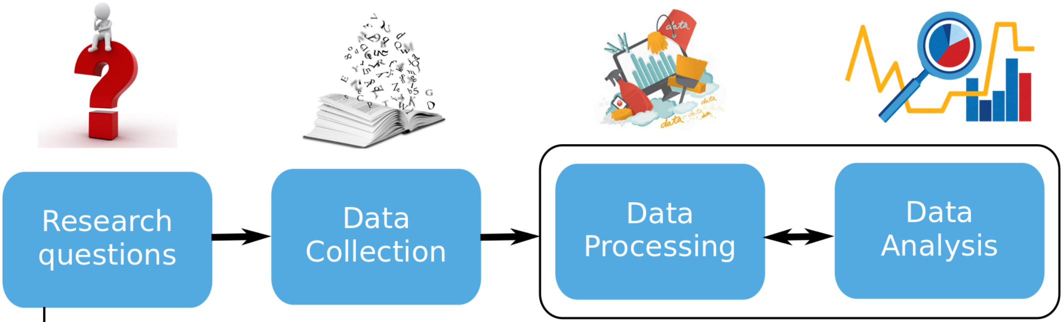
UF | IFAS Research UNIVERSITY of FLORIDA
UNIVERSITY OF LIVERPOOL
UNIVERSIDAD DE LIMA
HFSP

rocio.joo@ufl.edu
rocio_joo

Text analysis workflow



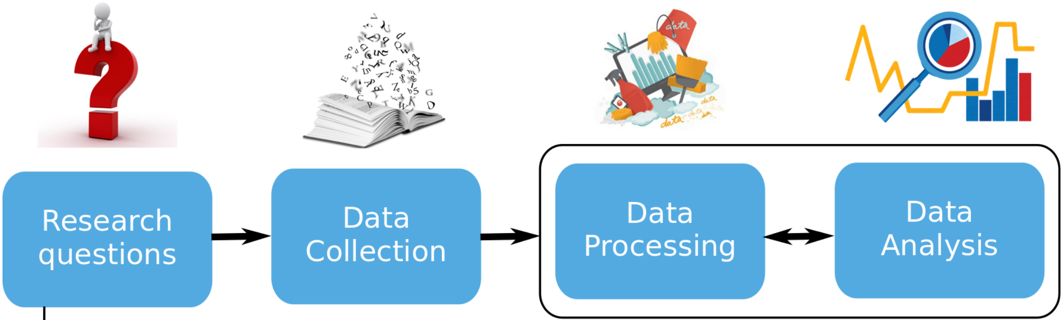
Text analysis workflow: movement ecology



- which topics?
- statistical methods?
- software?
- ⋮

in the last decade

Text analysis workflow: movement ecology



- which topics?
- statistical methods?
- software?
- ⋮

in the last decade

Text analysis workflow: movement ecology

Research questions → Data Collection → Data Processing ↔ Data Analysis

- which topics? **8007** abstracts
- statistical methods?
- software?
- ⋮

in the last decade

Text analysis workflow: movement ecology

Research questions → Data Collection → Data Processing ↔ Data Analysis

Research questions:
- which topics?
- statistical methods?
- software?
- ⋮

in the last decade

8007 abstracts

Data Processing:
- remove redundant words
- standard English (BrE or AmE)
- lemmatize
- filter out words appearing once

Text analysis workflow: movement ecology

Research questions → Data Collection → Data Processing ↔ Data Analysis

Research questions:
- which topics?
- statistical methods?
- software?
- ⋮

in the last decade

Data Collection: **8007** abstracts

Data Processing:
- remove redundant words
- standard English (BrE or AmE)
- lemmatize
- filter out words appearing once

Data Analysis → Topic Modeling

Topic modeling

Why attend IBC?

Why attend IBC?

Scientific community

- Community
- Network
- People
- Meeting
- Social

Knowledge

- Learn
- Teach
- Development
- Understand
- Study

Topic modeling

Why attend IBC?

Scientific community

- Community
- Network
- People
- Meeting
- Social

Knowledge

- Learn
- Teach
- Development
- Understand
- Study

"I can learn new things"

"It's nice to see the recent developments in the community"

"I am new in the field and want to meet people"

Why attend IBC?

Scientific community

- Community
- Network
- People
- Meeting
- Social

Knowledge

- Learn
- Teach
- Development
- Understand
- Study

"I can learn new things"

"It's nice to see the recent developments in the community"

"I am new in the field and want to meet people"

Topic modeling

Latent Dirichlet Allocation (LDA)

- Bayesian mixture model (Blei et al. 2003; Grün and Hornik 2011)

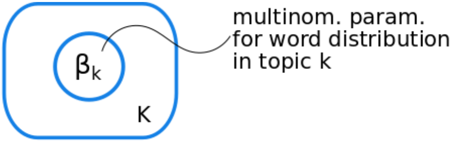
Latent Dirichlet Allocation (LDA)

- Bayesian mixture model (Blei et al. 2003; Grün and Hornik 2011)
- From a fixed number K of topics, each topic can be characterized by a multinomial distribution of words with parameter β, drawn from a Dirichlet distribution with param. δ

Topic modeling

Latent Dirichlet Allocation (LDA)

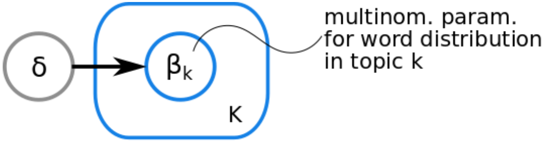
- Bayesian mixture model (Blei et al. 2003; Grün and Hornik 2011)
- From a fixed number K of topics, each topic can be characterized by a multinomial distribution of words with parameter β, drawn from a Dirichlet distribution with param. δ



Topic modeling

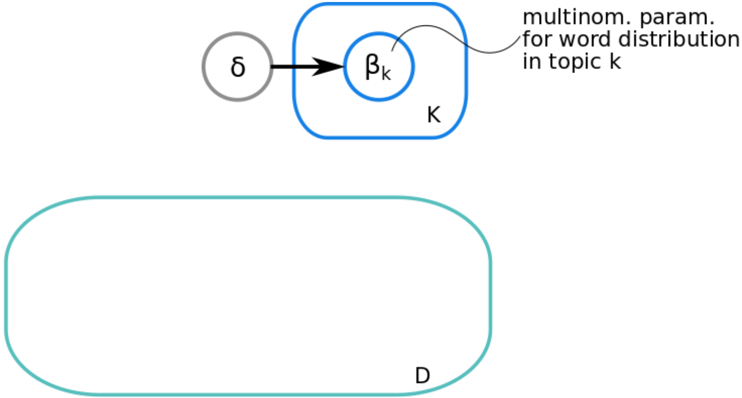
Latent Dirichlet Allocation (LDA)

- Bayesian mixture model (Blei et al. 2003; Grün and Hornik 2011)
- From a fixed number K of topics, each topic can be characterized by a multinomial distribution of words with parameter β, drawn from a Dirichlet distribution with param. δ



Latent Dirichlet Allocation (LDA)

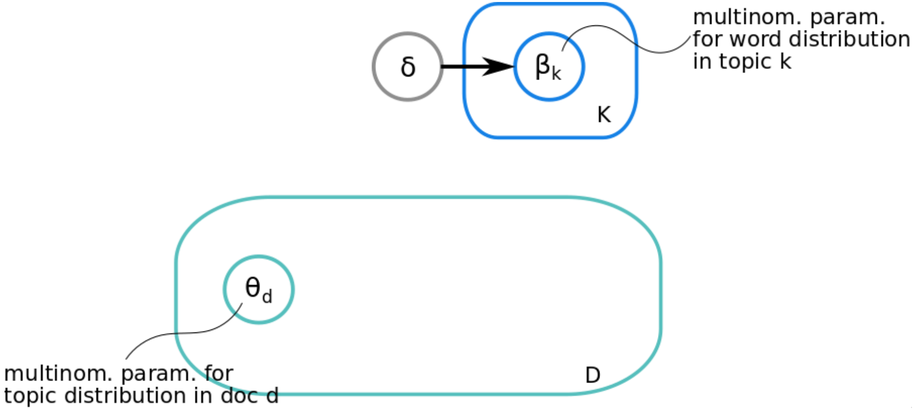
- Each document $\mathbf{d} \in \{1, ..., \mathbf{D}\}$ is composed by a mixture of topics, drawn from a multinomial distribution with parameter θ, which is drawn from a Dirichlet distribution with parameter α.



Topic modeling

Latent Dirichlet Allocation (LDA)

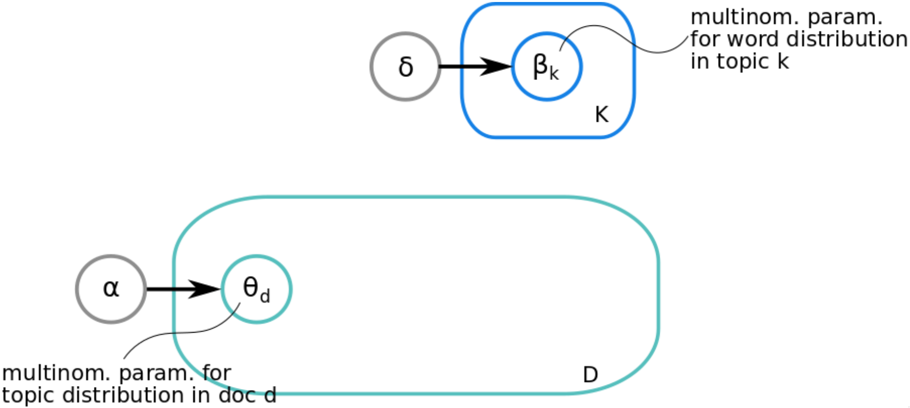
- Each document $\mathbf{d} \in \{1, ..., \mathbf{D}\}$ is composed by a mixture of topics, drawn from a multinomial distribution with parameter θ, which is drawn from a Dirichlet distribution with parameter α.



Topic modeling

Latent Dirichlet Allocation (LDA)

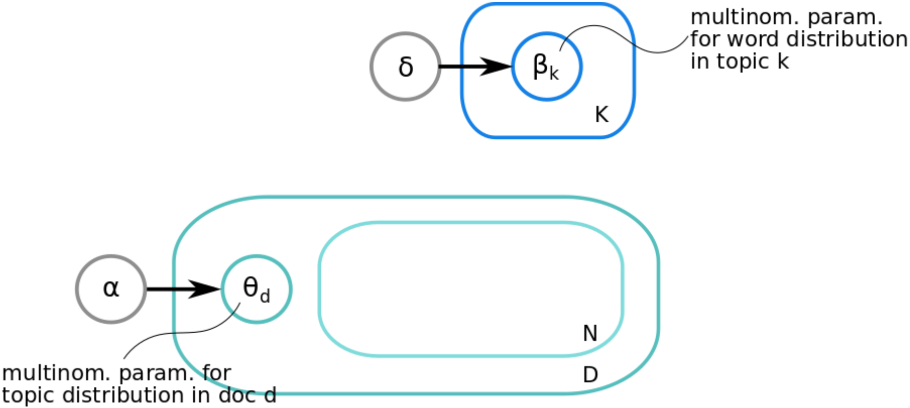
- Each document $\mathbf{d} \in \{1, ..., \mathbf{D}\}$ is composed by a mixture of topics, drawn from a multinomial distribution with parameter θ, which is drawn from a Dirichlet distribution with parameter α.



Topic modeling

Latent Dirichlet Allocation (LDA)

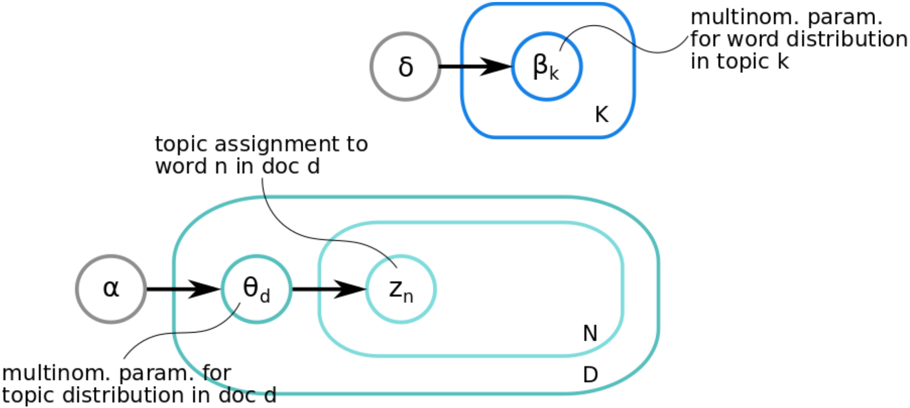
- For each word $\mathbf{w}$ in document $\mathbf{d}$, first a hidden topic $\mathbf{z}$ is selected from the multinomial distribution with parameter θ.



Topic modeling

Latent Dirichlet Allocation (LDA)

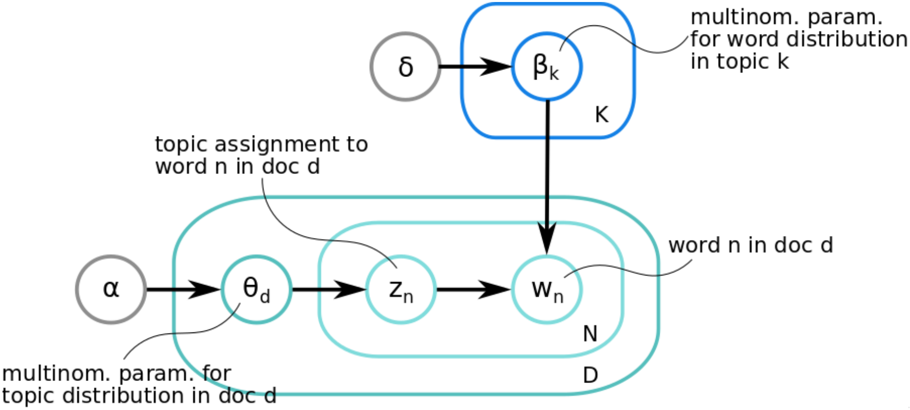
- For each word $\mathbf{w}$ in document $\mathbf{d}$, first a hidden topic $\mathbf{z}$ is selected from the multinomial distribution with parameter θ.



Topic modeling

Latent Dirichlet Allocation (LDA)

- For each word $\mathbf{w}$ in document $\mathbf{d}$, first a hidden topic $\mathbf{z}$ is selected from the multinomial distribution with parameter θ.
- From the selected topic $\mathbf{z}$, a word is selected based on the multinomial distribution with parameter β.

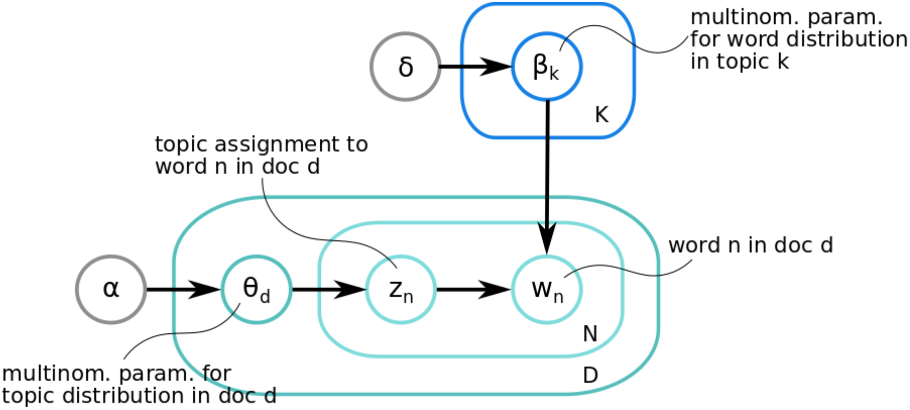


Topic modeling

Latent Dirichlet Allocation (LDA)

The log-likelihood of a document $\mathbf{d} = \{\mathbf{w}_1, ..., \mathbf{w}_N\}$ is

$$\mathbf{l}(\alpha, \beta) = \log(\mathbf{p}(\mathbf{d} \mid \alpha, \beta)) = \log \int \sum_{\mathbf{z}} \left[\prod_{\mathbf{n}=1}^{\mathbf{N}} \mathbf{p}(\mathbf{w}_\mathbf{i} \mid \mathbf{z}_\mathbf{i}, \beta)\mathbf{p}(\mathbf{z}_\mathbf{i} \mid \theta) \right] \mathbf{p}(\theta \mid \alpha) \mathbf{d}\theta$$



Topic modeling

Latent Dirichlet Allocation (LDA)

The log-likelihood of a document $\mathbf{d} = \{\mathbf{w}_1, ..., \mathbf{w}_N\}$ is

$$\mathbf{l}(\alpha, \beta) = \log(\mathbf{p}(\mathbf{d} \mid \alpha, \beta)) = \log \int \sum_{\mathbf{z}} \left[\prod_{\mathbf{n}=1}^{\mathbf{N}} \mathbf{p}(\mathbf{w_i} \mid \mathbf{z_i}, \beta)\mathbf{p}(\mathbf{z_i} \mid \theta) \right] \mathbf{p}(\theta \mid \alpha)\mathbf{d}\theta$$

Inference on the posterior via Variational Expectation Maximization (Blei et al. 2003; Grün and Hornik 2011); `topicmodels` R package

Assumptions:

- Exchangeability: order of words is negligible
- Topics are uncorrelated
- Number of topics is known (in this study: 15)

Topic modeling - LDA

Results - movement ecology

Topic modeling - LDA

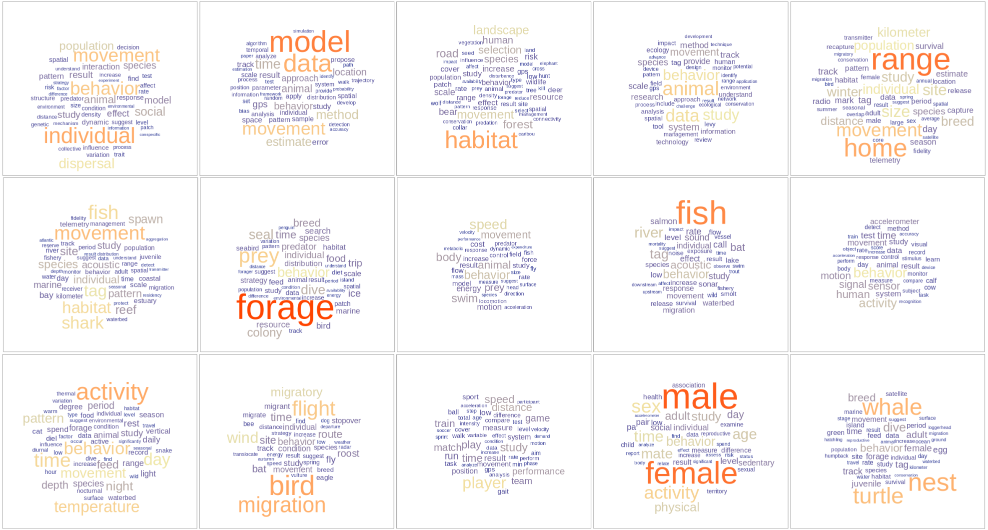
Results - movement ecology

- $\mathbf{E}(\beta \mid \mathbf{z}, \mathbf{w}) \longrightarrow$ word distribution per topic $\longrightarrow$ label topics

Topic modeling - LDA

Results - movement ecology

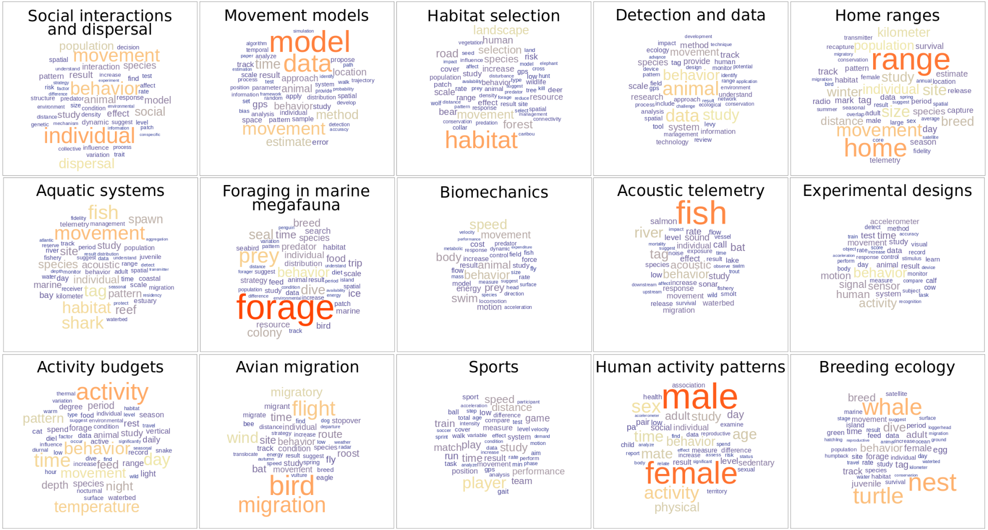
- $\mathbf{E}(\beta \mid \mathbf{z}, \mathbf{w}) \rightarrow$ word distribution per topic $\rightarrow$ label topics



Topic modeling - LDA

Results - movement ecology

- $\mathbf{E}(\beta \mid \mathbf{z}, \mathbf{w})$ $\rightarrow$ word distribution per topic $\rightarrow$ label topics



Topic modeling - LDA

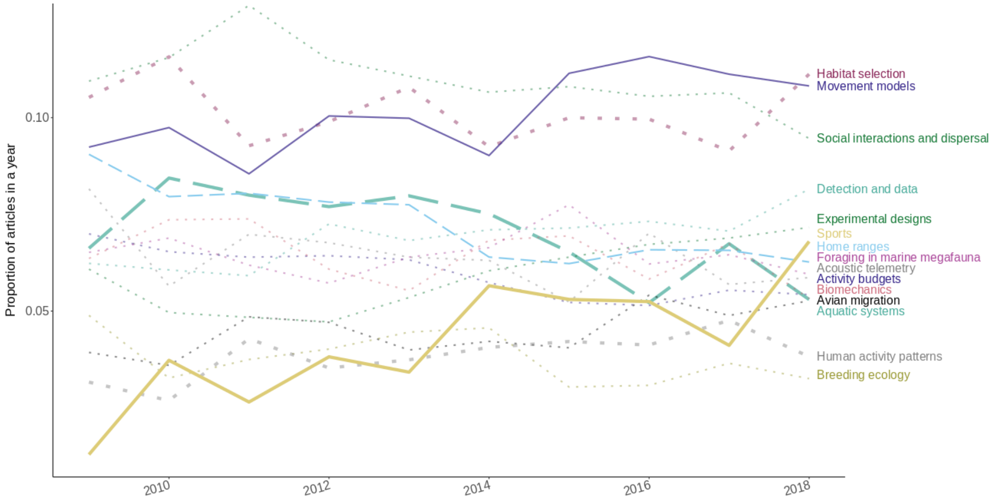
Results - movement ecology

- $\mathbf{E}(\theta_{\mathbf{d}} \mid \mathbf{z})$: topic distribution per document
- $\sum_{\mathbf{d}} \mathbf{E}(\theta_{\mathbf{d}} \mid \mathbf{z_k})$: proxy of prevalence of each topic

Topic modeling - LDA

Results - movement ecology

- $\mathbf{E}(\theta_\mathbf{d} \mid \mathbf{z})$: topic distribution per document
- $\sum_\mathbf{d} \mathbf{E}(\theta_\mathbf{d} \mid \mathbf{z_k})$: proxy of prevalence of each topic



Further exploration?

- Topic evaluation: word intrusion
  - Take the highest probability words from a topic
  - Take a high-probability word from another topic and add it
  - Ask humans to identify the word that does not belong

Results shown here: Joo et al. pre-print.

References

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. ‘Latent Dirichlet Allocation.’ Journal of Machine Learning Research 3 (Jan): 993âĂŞ1022. http://jmlr.csail.mit.edu/papers/v3/blei03a.html.
- Grün, Bettina, and Kurt Hornik. 2011. ‘topicmodels: An R Package for Fitting Topic Models.’ Journal of Statistical Software 40 (13): 1âĂŞ30. https://doi.org/10.18637/jss.v040.i13.
- Joo, Rocío, Simona Picardi, Matthew E. Boone, Thomas A. Clay, Samantha C. Patrick, Vilma S. Romero-Romero and Mathieu Basille. Pre-print. ‘A decade of movement ecology’. arxiv. https://arxiv.org/abs/2006.00110.

Thanks for your attention

rocio.joo@ufl.edu

rocio_joo