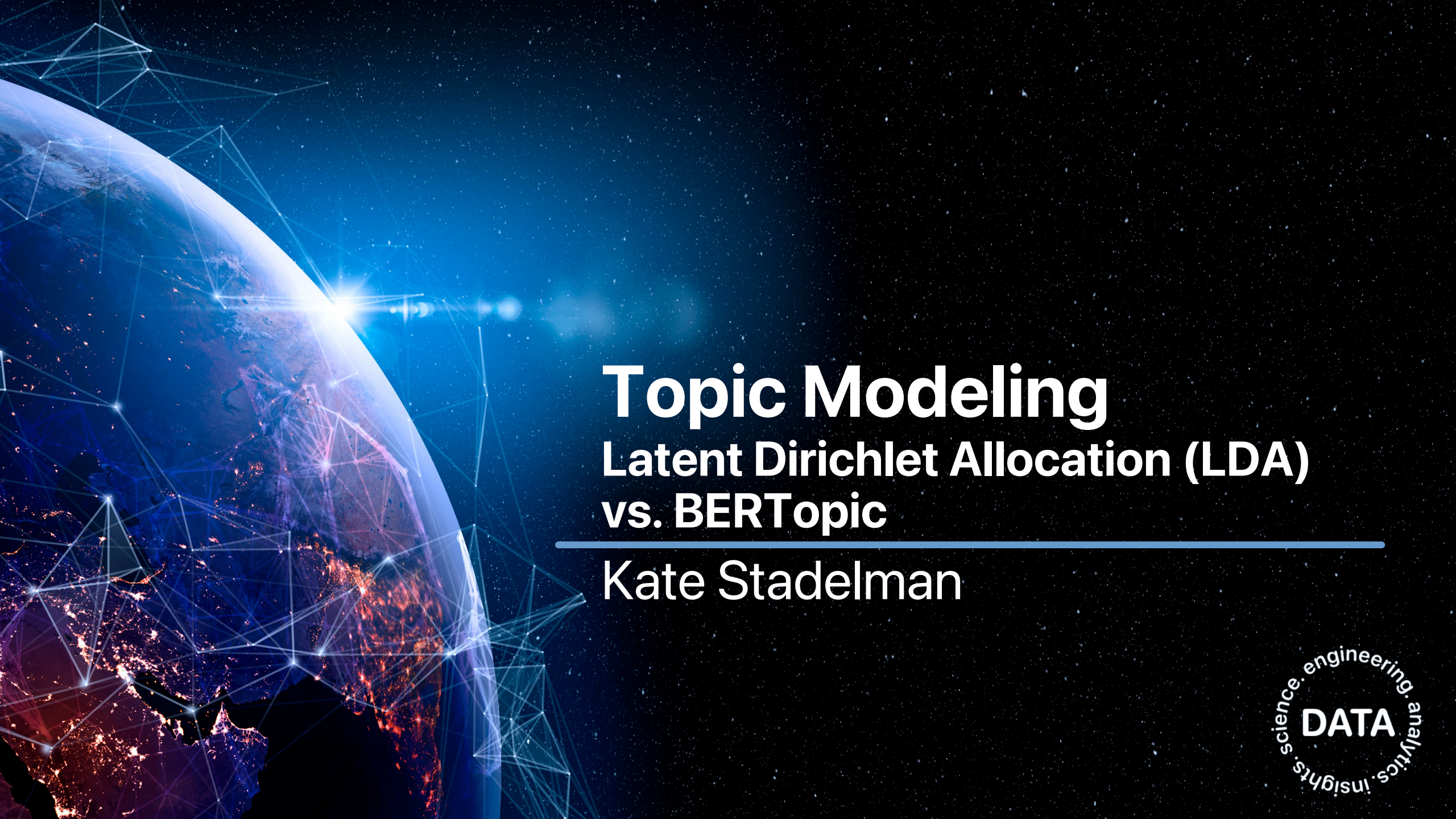

Topic Modeling

Latent Dirichlet Allocation (LDA) vs. BERTopic

Kate Stadelman

Disclaimer

The opinions expressed here are solely my own and do not reflect the views or opinions of my employer.

All code in this project is derived from publicly available open-source libraries and not considered proprietary.

About Me

Core Competencies

- Data Engineering & Warehousing
- Healthcare Information
- Business Intelligence Reporting
- Advanced Analytics & Data Science

Education

- MS in Data Science, 2022
- BS in Mathematics, 2006

My LinkedIn

My Dogs

We Can Ride

Topic Modeling GitHub Repo

All code from today's talk is available here:

kaspii314/topic_modeling: Overview of Topic Modeling using Latent Dirichlet Allocation and BERTopic. (github.com)

Talk Overview

- What is Topic Modeling?
- Latent Dirichlet Allocation (LDA) vs. BERTopic
- Review Key NLP Concepts
- Exploratory Data Analysis
- LDA Topic Model
- BERTopic Model

What is Topic Modeling?

Topic modeling is a Natural Language Processing (NLP) technique for discovering topics in a collection of documents, allowing you to see hidden structure in your text data.

Latent Dirichlet Allocation (LDA) (2003)

- David Blei, Andrew Ng and Michael I. Jordan

BERTopic (2022)

- Maarten Grootendorst

LDA vs. BERTopic

Latent Dirichlet Allocation (LDA)	BERTopic
Bayesian probabilistic model that estimates probability distributions for topics in documents and words in topics.	Algorithm that leverages the transformer-based model BERT and c-TF-IDF to "create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions." (Grootendorst, 2022)
Each document is assigned topic probabilities. - Document 1 is 80% Topic 3, 15% Topic 2, 5% Topic 5	Each document is assigned to a single topic. - Document 1 is Topic 3
While extracted topics can be used to train classification models, topic generation is unsupervised.	Topic generation is unsupervised by default, but semi-supervised and supervised learning approaches are supported.

LDA vs. BERTopic (cont)

Latent Dirichlet Allocation (LDA)	BERTopic
All documents are assigned to at least one topic.	Outlier documents are assigned to Topic -1 (no real topic).
No simple mechanism to merge topics.	Overlapping topics can be easily merged.
Coherence and perplexity measures readily available for topic evaluation. In practice, these measures may be less effective than using visualizations.	While you can maneuver Gensim (LDA) to calculate coherence and perplexity scores on BERTopic topics, BERTopic does not have these built-in for topic evaluation. However, there are many excellent visualizations available.

Key NLP Concepts

- Tokenization
- Stemming
- Lemmatization
- Stop Words
- TF-IDF

Exploratory Data Analysis
LDA Model
BERTopic Model

Jupyter Notebook

Further Reading & Resources

LDA Topic Modeling

- Topic Models: Past, Present, Future
- Gensim: Topic Modeling for Humans
- A Beginner's Guide to Latent Dirichlet Allocation (LDA)
- Evaluate Topic Models: Latent Dirichlet Allocation (LDA)

BERTopic

- BERTopic
- Advanced Topic Modeling with BERTopic
- BERTopic Explained

References

Harshal H. (2023, September). *Starbucks Reviews Dataset* [Data set]. Kaggle. Retrieved October 22, 2023 from https://www.kaggle.com/datasets/harshalhonde/starbucks-reviews-dataset/

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*. https://arxiv.org/abs/2203.05794

Rehurek, R. and Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. *Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks, Valletta, Malta*, 46-50. https://radimrehurek.com/lrec2010_final.pdf