syed_2017_full_text_or_abstract_examining_topic_coherence_scores_using_latent_dirichlet_allocation

Year

2017

Author(s)

Syed, Shaheen and Spruit, Marco

Title

Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation

Venue

DSAA

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling assisted by associated documents (titles and contents)

Topic labeling parameters

Nr of inspected terms: 15

Label generation

A fisheries domain expert was available to manually label and rank the topics as an alternative means of assessing the quality of topics.

The LDA model with the optimal coherence score, obtained with an elbow method (the point with maximum absolute second derivative), was analyzed by a fisheries domain expert.
The domain expert is affiliated with the leading competence institution for fishery and aquaculture in Norway.

The analysis consisted of an inspection of the top 15 most probable words for each topic, together with an inspection of the document titles and content.
The domain expert attached a label to each topic that best captured the semantics of the top 15 words.

TABLE IV
A SELECTION OF TOPICS FROM DS_1 WITH THE 15 MOST PROBABLE WORDS, TOPIC LABEL, AND RANKING DATA. TEXT IN BOLD INDICATES INCORRECT TERMS.

Dataset	Label	Top 15 words	Ranking
Abstract	fish distribution	fishing, distribution, data, species, areas, catch, abundance, spatial, habitat, model, fishery, effort, fish, water, sea	High
	Population models	model, data, mortality, stock, fish, population, fishing, models, recruitment, cod, estimates, **using**, size, rates, **used**	Medium
	Population genetics	genetic, populations, population, **among**, lake, fish, loci, microsatellite, **two**, structure, diversity, **within**, samples, species, river	Low
Full-text	Salmon population dynamics	salmon, trout, prey, growth, atlantic, temperature, water, rate, juvenile, salmo, feeding, wild, density, food, populations	High
	Population genetics	genetic, populations, population, river, samples, loci, salmon, **among**, dna, atlantic, sample, sea, microsatellite, structure, alleles	Medium

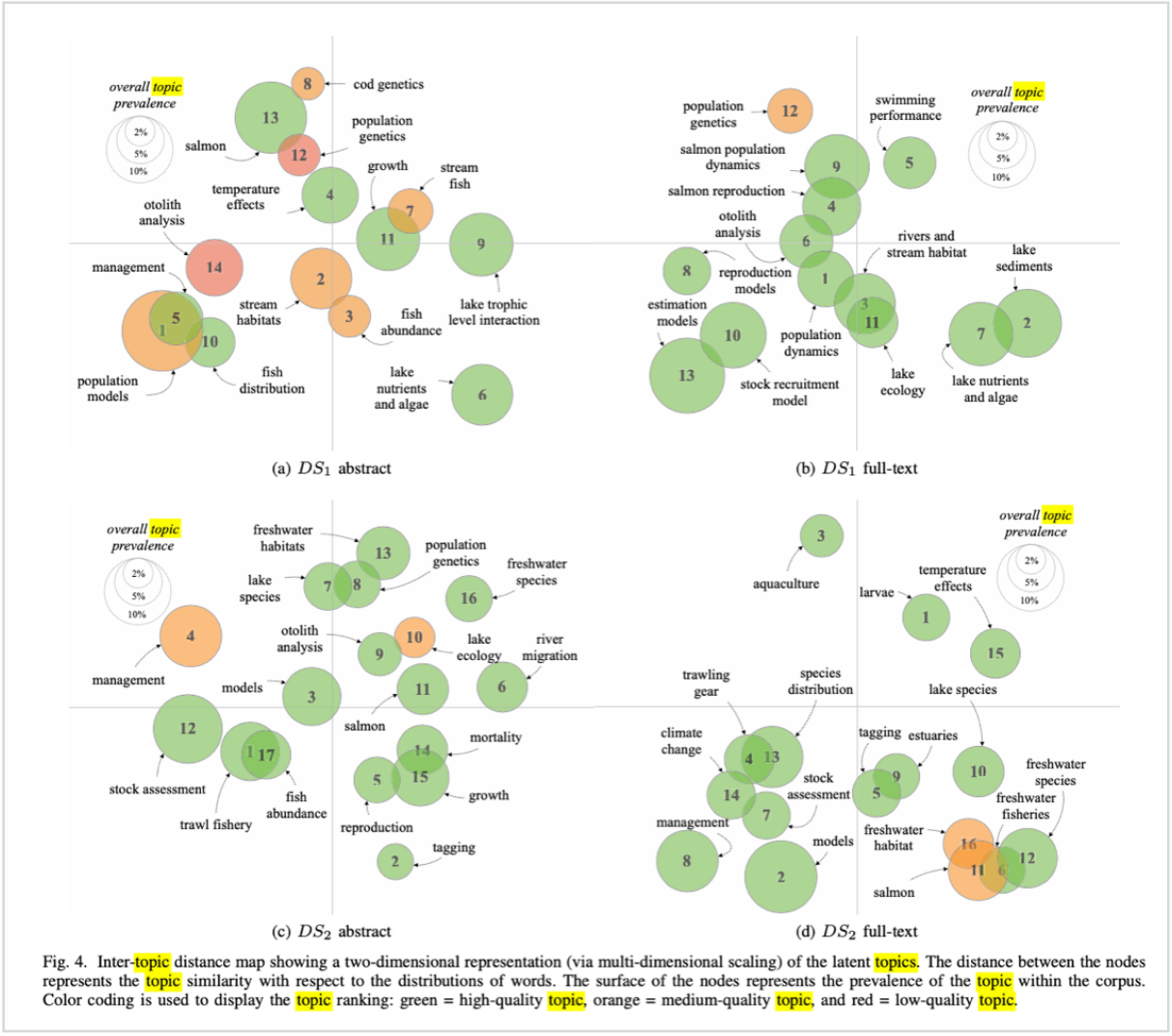

Fig. 4. Inter-topic distance map showing a two-dimensional representation (via multi-dimensional scaling) of the latent topics. The distance between the nodes represents the topic similarity with respect to the distributions of words. The surface of the nodes represents the prevalence of the topic within the corpus. Color coding is used to display the topic ranking: green = high-quality topic, orange = medium-quality topic, and red = low-quality topic.

Motivation

Besides topic coherence, we explore the effects of human topic ranking– often considered the gold standard for topic interpretability– on topics uncovered from abstract and full-text data.

Topic modeling

LDA

Topic modeling parameters

Nr of topics (K): 1 to 40

smoothing of words within topics η : 1 / V

Topics within documents α : 1 / K

Convergence iteration parameter for the expectation step (E-step): 100

Nr. of topics

13 (DS1), 16 (DS2)

Label

Single or multi word label manually assigned by domain expert

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Topic modeling evaluation
Dataset: Fisheries and Aquatic Science

Problem statement

This paper assesses topic coherence and human topic ranking of uncovered latent topics from scientific publications when utilizing the topic model latent Dirichlet allocation (LDA) on abstract and full-text data.
The coherence of a topic, used as a proxy for topic quality, is based on the distributional hypothesis that states that words with similar meaning tend to co-occur within a similar context.
Little is known about the effects of different types of textual data on generated topics. Our research is the first to explore these practical effects and shows that document frequency, document word length, and vocabulary size have mixed practical effects on topic coherence and human topic ranking of LDA topics.

We furthermore show that large document collections are less affected by incorrect or noise terms being part of the topic-word distributions, causing topics to be more coherent and ranked higher.

Corpus

DS1

Origin: Canadian Journal of Fisheries and Aquatic Science
Nr. of documents: 4,417
Details:

- Research articles (1996 to 2016)
- studies where a single scientific journal was analyzed from a domain-specific journal

DS2

Origin: 12 top-tier fisheries journals
Nr. of documents: 15,004
Details:

- esearch articles (2000 to 2016)
- studies where LDA was used to uncover topics from a multitude of related domain-specific journals

Document

Research article abstract and full- text data

Pre-processing

- Converted from PDF to plain text
- Tokenized
- Single-character words, numbers, and punctuation marks were removed
- Removed all single-occurrence words, words that occurred in more than 90% of the documents, and words that belonged to a standard English stop word list (n = 153).

```
@inproceedings{syed_2017_full_text_or_abstract_examining_topic_coherence_scores
_using_latent_dirichlet_allocation,
  author = {Syed, Shaheen and Spruit, Marco},
  booktitle = {2017 IEEE International Conference on Data Science and Advanced
```

```
Analytics (DSAA)},
  date-added = {2023-04-05 12:31:09 +0200},
  date-modified = {2023-04-05 12:31:09 +0200},
  doi = {10.1109/DSAA.2017.61},
  pages = {165-174},
  title = {Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation},
  year = {2017}}
```

#Thesis/Papers/BS