20_12_2022

Thesis title

"Topic labeling in the period 2017-2022: A Systematic Literature Review"
"A Systematic Literature Review on topic labeling research in the period 2017-2022"

EMNLP 2022

From the 2022 instance of the EMNLP conference which took place from December 7th to the 11th a set of 828 papers is collected.

To this initial set of papers, the previously presented query `"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))` is applied, resulting in a subset of 20 articles (29 without the proximity operator).

After the application of the agreed upon inclusion and exclusion criteria, no paper is deemed relevant to the current SLR and added to the original set.

Paper graph

litmaps.com
Using this tool in addition to only relying on local sources could allow us to:

- Capture richer information about the interested papers (for example, by also highlighting connections between references/citations).
- Generate easily shareable interactive graphs.

Caveats:

1. Generating multiple maps, or maps larger than 100 nodes requires a Pro account. I have written to the support team to ask what happens to generated maps after an account loses the Pro status.
2. Some very recent papers are not yet available on the platform (this should be solved on the next DB refresh).

Currently missing papers

- One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality
- Community Topic: Topic Model Inference by Consecutive Word Community Discovery
- Improving Deep Embedded Clustering via Learning Cluster-level Representations

Non-retrievable Forward Snowballing papers

There is currently a set of 13 papers (out of the total 1147) that are not currently retrievable (neither through the university, nor through Sci-Hub)
Retrieval requests have been sent to the library.

Imposing a quality constraint on papers retrieved from forward snowballing?

An initial note on the data collection process

During the data collection process, when deciding upon the data items that will need to be extracted from the selected papers, a distinction will probably need to be made on the analysis performed between papers employing manual labeling and the other encountered labeling techniques.
As expected, papers making use of manual topic labeling techniques represent the majority of the selected work.
Despite this fact, these papers are generally less interesting for the purposes of our research and probably do not require a very detailed exploration when it comes to the data collection phase.
The data gathering process should therefore be structured in such a way as to account for this fact.

Forward snowballing (Incoming citations)

After having obtained the relevant (outgoing) references from the selected set of papers following the backward snowballing activity described in the previous section, the similar process of forward snowballing was carried out in order to to capture the relevant (incoming) citations.
The process follows a similar set of steps as the ones described for backward snowballing, with the two major differences being:
- The fact that an external repository (SemanticScholar) is used to obtain the citations since, unlike references, they cannot be extracted directly from the paper that is being examined.
- No filtering step based on publication time needs to be applied (since all citation will inherently respect the imposed time constraints).

Initial results

Executing forward snowballing on the set of 65 initially selected papers resulted in **1161** extracted citations. Once again, the query (and proximity operator) are applied to the extracted citation. This generates a set of **358** items (590 without the proximity operator) to which the inclusion/exclusion criteria need to be applied to.

B.S. selection (Analysed 203 out of 358, Selected 44 out of 203)

- A Computational Analysis of News Media Bias: A South African Case Study
- A high-resolution temporal and geospatial content analysis of Twitter posts related to the COVID-19 pandemic
- A Semi-automated Approach for Identification of Trends in Android Ransomware Literature
- A Structural Topic Modeling-Based Bibliometric Study of Sentiment Analysis Literature
- A topic-based patent analytics approach for exploring technological trends in smart manufacturing
- An Analysis of Programming Course Evaluations Before and After the Introduction of an Autograder
- Analyzing genderless fashion trends of consumers' perceptions on social media: using unstructured big data analysis through Latent Dirichlet Allocation-based topic modeling
- Automatic Labeling of Topic Models Using Graph-Based Ranking
- Automatic Topic Labeling model with Paired- Attention based on Pre-trained Deep Neural Network
- Automatic topic labeling using graph-based pre-trained neural embedding
- Customer expectations in the hotel industry during the COVID-19 pandemic: a global perspective using sentiment analysis
- Discovering Interpretable Topics by Leveraging Common Sense Knowledge
- Exploring Sources of Satisfaction and Dissatisfaction in Airbnb Accommodation Using Unsupervised and Supervised Topic Modeling
- Exploring the Development of Research, Technology and Business of Machine Tool Domain in New-Generation Information Technology Environment Based on Machine Learning
- Exploring the links between research demand and supply: The case of Chagas
- Exposing Emerging Trends in Smart Sustainable City Research Using Deep Autoencoders-Based Fuzzy C-Means
- Extractive and Abstractive Sentence Labelling of Sentiment-bearing Topics
- Finding Scientific Topics in Continuously Growing Text Corpora
- Harnessing the "Wisdom of Employees" from Online Reviews
- Hierarchical Interpretation of Neural Text Classification
- Identification of topic evolution: network analytics with piecewise linear representation and word embedding
- Identifying Major Research Areas and Minor Research Themes of Android Malware Analysis and Detection Field Using LSA
- Job satisfaction and employee turnover determinants in high contact services:

Insights from Employees'Online reviews

- Knowledge Source Rankings for Semi-Supervised Topic Modeling
- Mining quality determinants of product-service systems from user-generated contents
- Modeling the public attitude towards organic foods: a big data and text mining approach
- Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents
- On-Demand Recent Personal Tweets Summarization on Mobile Devices
- Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling
- Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews
- Recent trends of green human resource management: Text mining and network analysis
- Revealing industry challenge and business response to Covid-19: a text mining approach
- TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters
- The informational value of employee online reviews
- The Legitimacy of Wind Power in Germany
- The Performance of Topic Evolution Based on a Feature Maximization Measurement for the Linguistics Domain
- The use of citation context to detect the evolution of research topics: a large-scale analysis
- The Voice of Drug Consumers: Online Textual Review Analysis Using Structural Topic Model
- TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition
- Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations
- Topic evolution, disruption and resilience in early COVID-19 research
- Topic modelling for theme park online reviews: analysis of Disneyland
- Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis
- Understanding Airline Passengers during Covid-19 Outbreak to Improve Service Quality: Topic Modeling Approach to Complaints with Latent Dirichlet Allocation Algorithm
- Understanding Anonymous Social Media Posts using Topic Modeling
- Understanding Research Trends in Android Malware Research Using Information

Modelling Techniques

- Value creation in emerging technologies through text mining: the case of blockchain
- What Do Websites Say about Internet of Things Challenges? A Text Mining Approach
- What matters most to patients? On the Core Determinants of Patient Experience from Free Text Feedback
- What we talk about when we talk about EEMs: using text mining and topic modeling to understand building energy efficiency measures (1836-RP)
- Revealing industry challenge and business response to Covid-19: a text mining approach
- The use of citation context to detect the evolution of research topics: a large-scale analysis
- A Structural Topic Modeling-Based Bibliometric Study of Sentiment Analysis Literature
- Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis
- Exposing Emerging Trends in Smart Sustainable City Research Using Deep Autoencoders-Based Fuzzy C-Means
- Exploring Sources of Satisfaction and Dissatisfaction in Airbnb Accommodation Using Unsupervised and Supervised Topic Modeling
- Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews
- A Computational Analysis of News Media Bias: A South African Case Study
- Topic evolution, disruption and resilience in early COVID-19 research
- A high-resolution temporal and geospatial content analysis of Twitter posts related to the COVID-19 pandemic
- An Analysis of Programming Course Evaluations Before and After the Introduction of an Autograder
- Understanding Anonymous Social Media Posts using Topic Modeling
- Customer expectations in the hotel industry during the COVID-19 pandemic: a global perspective using sentiment analysis
- Exploring the Development of Research, Technology and Business of Machine Tool Domain in New-Generation Information Technology Environment Based on Machine Learning
- Value creation in emerging technologies through text mining: the case of blockchain
- Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education
- Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation
- Understanding the temporal evolution of COVID-19 research through machine

learning and natural language processing
- Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling
- A Decade of Sentic Computing: Topic Modeling and Bibliometric Analysis
- Selecting the appropriate leading journal in Hospitality and Tourism research: a guide based on the topic-journal fit and the JCR impact factor
- Open data visualizations and analytics as tools for policy-making
- Exploring the nexus between food and veg*n lifestyle via text mining-based online community analytics
- Listen to E-scooter riders: Mining rider satisfaction factors from app store reviews
- On the evolution of research in hypersonics: application of natural language processing and machine learning

Additionally, the following 13 papers that were already part of the initial selection of publications were retrieved once again following the forward snowballing procedure:
- Automatic Generation of Topic Labels
- One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality
- Multilingual Topic Labelling of News Topics using Ontological Mapping
- Multimodal Topic Labelling
- BART-TL: Weakly-Supervised Topic Label Generation
- Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling
- Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents
- Automatic topic labeling using graph-based pre-trained neural embedding
- A Semi-automated Approach for Identification of Trends in Android Ransomware Literature
- Job satisfaction and employee turnover determinants in high contact services: Insights from Employees'Online reviews
- Harnessing the "Wisdom of Employees" from Online Reviews
- Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures
- The informational value of employee online reviews
- Understanding Research Trends in Android Malware Research Using Information Modelling Techniques
- Exploring the links between research demand and supply: The case of Chagas
- A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews
- Finding Scientific Topics in Continuously Growing Text Corpora

- Community Topic: Topic model inference by consecutive word community discovery
- What Do Websites Say about Internet of Things Challenges? A Text Mining Approach
- Neural Topic Modeling with Cycle-Consistent Adversarial Training
- A weakly-supervised graph-based joint sentiment topic model for multi-topic sentiment analysis
- Influence Factorization for identifying authorities in Twitter
- Supporting digital content marketing and messaging through topic modelling and decision trees
- Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response

Finally, one (1) papers had already been selected during the backwards snowballing phase:
- Transfer Topic Labeling with Domain-Specific Knowledge Base: An Analysis of UK House of Commons Speeches 1935–2014

F.S. venues (ongoing)
The following is the list of previously unexplored venues to which the work selected by the forward snowballing task belongs to:

Conferences
- Asia-Pacific Chapter of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (AACL-IJCNLP)
- Annual Workshop of the Australasian Language Technology Association
- Conference on Information Technology Based Higher Education and Training (ITHET)
- International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)
- International Conference on Information Systems (ICIS)
- International Joint Conference on Neural Networks (IJCNN)
- South African Institute of Computer Scientists and Information Technologists (SAICSIT)
- Conference on Information Technology Based Higher Education and Training (ITHET)

Journals
- Annals of Tourism Research
- Cognitive Computation
- Complexity
- Computers & Education
- European Journal of Operational Research
- Fashion and Textiles

- Frontiers in Psychology
- IEEE Access
- Information
- International Journal of Contemporary Hospitality Management
- International Journal of Environmental Research and Public Health
- International Journal of Hospitality Management
- Journal of Big Data
- Journal of Computational Social Science
- Journal of Manufacturing Technology Management
- Journal of Medical Internet Research
- Journal of the Association for Information Science and Technology
- Journal of Travel & Tourism Marketing
- Neurocomputing
- Quality Engineering
- Scholarly Document Processing (SDP)
- Science and Technology for the Built Environment
- Scientometrics
- Sustainability
- Technology Analysis & Strategic Management
- Tourism Recreation Research
- Transportation Research Record
- Food Quality and Preference
- Government Information Quarterly

Next steps
- Finish forward snowballing selection
- Finish working on paper graphs
- **Transcribe the work done so far on Overleaf**
- Establish details of data collection process

#Thesis/Weekly notes#