07_12_2022

Grey literature

“Information produced on all levels of government, academia, business and industry in electronic and print formats not controlled by commercial publishing” ie. where publishing is not the primary activity of the producing body.” - Third International Conference on Grey Literature in 1997

“A Multivocal Literature Review (MLR) is a form of a Systematic Literature Review (SLR) which includes the grey literature (e.g., blog posts and white papers) in addition to the published (formal) literature (e.g., journal and conference papers).” - Guidelines for including grey literature and conducting multivocal literature reviews in software engineering

Initial definition of Topic Label(ing)

In order to find an early definition of Topic label(ing), the Semantic Scholar repository was queried using the keyword `"topic label*"` and sorting the results by citation count. Additionally, a date range restricting the results to paper released after 2003 was also set. This constraint was imposed due to the fact that the first application of Latent Dirichtlet Allocation in machine learning (Blei et al., 2003) was published in 2003.

“Automatic Labeling of Multinomial Topic Models” - Mei et al., 2007
“(**Topic Model**) A topic model θ in a text collection C is a probability distribution of words {p(w|θ)} w∊V where V is a vocabulary set. […] Intuitively, a topic model can represent a semantically coherent topic in the sense that the high probability words often collectively suggest some semantic theme.
[…]
(**Topic Label**) A topic label, or a “label”, l, for a topic model θ, is a sequence of words which is semantically meaningful and covers the latent meaning of θ.
Words, phrases, and sentences are all valid labels under this definition.”

Initial guidelines for paper graph generation

- Highlight in different colours papers from the original selection, the BS set and the FS set.
- For the graph, include ALL references (not just the ones published after 2017)
  - Highlight differently the ones belonging to the chosen time period (2017 onwards).

- By keeping older references, we might identify topically relevant research that might be foundational for the work we do analyse.
    - Example: "Automatic labelling of topics with neural embeddings" - Bhatia et al., 2016

Non-retrievable Forward Snowballing papers

There is currently a set of 13 papers (out of the total 1161) that is not currently retrievable (neither through the university, nor through Sci-Hub)

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 **1147** 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.

F.S. selection (ongoing, currently analysed 125 out of 358, selected 36 out of 125)

- 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)

Additionally, the following 9 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

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)

Journals

- Annals of Tourism Research
- Cognitive Computation
- Complexity
- 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

Next steps

- Generate the final selection for forward snowballing
- Depending on the size of the final selection (after snowballing)
    - Increase/Decrease the time-frame / selected venues
- Build paper graphs using Pajek
- Start to rewrite the notes into the “Methods” section (i.e. transcribe the work done so far on Overleaf)
- **After December 11th, gather papers from** emnlp 2022
- Establish details of data collection process

#Thesis/Weekly notes#