Directed topic extraction with side information.

Maria Osipenko[1]

Statistische Woche, TU Dortmund

[1]Hochschule für Wirtschaft und Recht Berlin; osipenko@hwr-berlin.de

Motivation

- Growing interest to sustainable investments

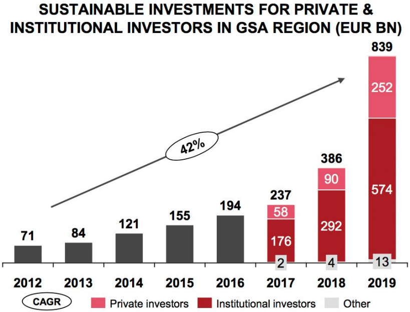


Figure 1: Source: Consultancy.eu

- Investment decisions integrate individual value systems
- Aligning investments with individual preferences
  - how to quantify sustainability?
  - how to compare investment possibilities?

Motivation

- Environment, social, governance (ESG) ratings diverge:

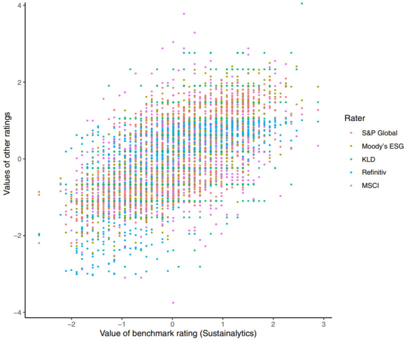


Figure 2: Ratings of different providers against a benchmark. Source: Berg, Kölbel, and Rigobon (2022) "Aggregated confusion The Divergence of ESG Ratings"

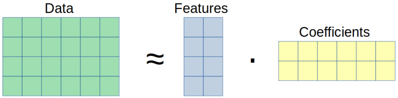
Motivation

- Kang and Kim (2022): Another source of information easily available to private investors
  - corporate responsibility reports
  - sustainability reports
  - environmental action reports
- A systematics e.g. in commonly accepted 17 UN sustainable development goals (SDGs) is at hand.

→ leverage information from these sources via automatic topic extraction while considering the value system established by the 17 SDGs.

Methods available

- Topic analysis: represent each document/ context in a low dimensional latent topic space:
  - Specific for topic extraction: Latent (probabilistic) Semantic Analysis, Latent Dirichlet allocation (LDA) and extentions thereof.
  - General purpose matrix factorization (MF) methods: Principal component analysis, Non-negativ matrix factorization, probabilistic versions and extensions thereof.



Methods available

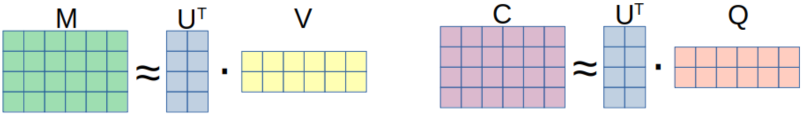
- How to embed known structure or side information in the unsupervised techniques?
  - keyword seeded LDA: Watanabe and Zhou (2022) and Eshima, Imai, and Sasaki (2023)
  - graph regularized MF: Rao et al. (2015) and Zhang et al. (2020) (recommendations)
  - common subspace projection/ subspace alignment (Fernando et al. (2013) for domain adaptation)
  - matrix co-factorization (MCF) techniques: Fang and Si (2011) (user communities) and Luo et al. (2019) (recommender systems)

$\rightarrow$ adopt MCF for topic extraction with side information.

Our approach

Decompose two term-context matrices (M from the sustainability reports and C from the SDG texts) jointly.

$$M \approx U^\top V \text{ and } C \approx U^\top Q$$



- M is the (weighted) term-context matrix for the corporate reports with dimensions $(p \times n)$, where p is the joint vocabulary.
- C is the (weighted) term-context matrix for the sustainability goals with dimensions $(p \times m)$, where p is again the joint vocabulary.
- U is the term-topic representation matrix of dimensions $(p \times k)$, where k is the number of topics.
- V/Q is the context-topic representation matrix for the reports/SGDs of dimensions $(k \times n)$.

Our approach

The associated MCF problem is then:

$$\min(||M - U^\top V||^2 + \lambda||C - U^\top Q||^2)$$

where λ adapts the importance of the loss on the second factorization term.

To preserve the non-negativity of the entries in M and C, to inhance interpretability $\rightarrow$ restrict the components to be non-negative:

$$\text{s.t. } U, V, Q \geq 0 \text{ elementwise.}$$

The algorithm

- alternating minimization/ alternating projection
- hierarchical non-negative alternating least squares (HALS) of Cichocki, Zdunek, and Amari (2007)
- with a modification for the co-factorization setup

Algorithm 1 HALS algorithm for MCF

---

while not converged **do**
  for $k = 1$ to K **do**
    update $V_k \leftarrow \max\left(\frac{U_k(M - U_{-k}^\top V_{-k})}{U_k U_k^\top}, 0\right)$
    update $Q_k \leftarrow \max\left(\frac{U_k(C - U_{-k}^\top Q_{-k})}{U_k U_k^\top}, 0\right)$
    update $U_k^\top \leftarrow \max\left(\frac{(M - U_{-k}^\top V_{-k})V_k^\top + \lambda(C - U_{-k}^\top Q_{-k})Q_k^\top}{V_k^\top V_k + \lambda Q_k^\top Q_k}, 0\right)$
  end for
end while

X_k denotes the kth row of the matrix X and X_{-k} denotes the matrix without its kth row.

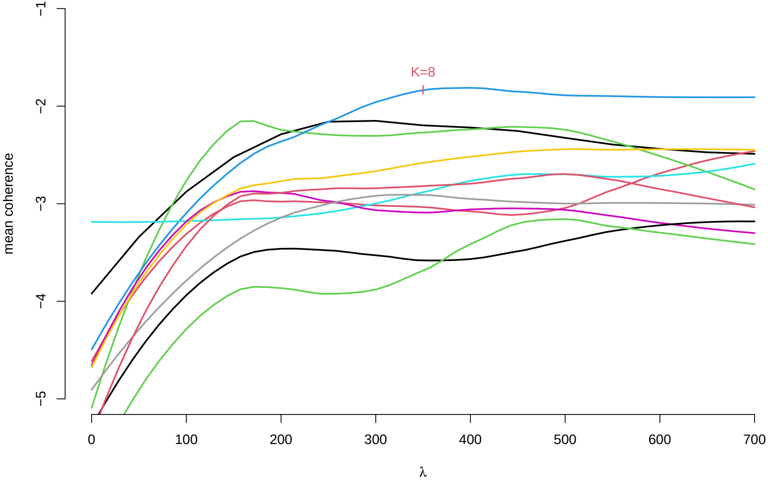
The data

- Corporate responsibility/sustainability reports: AAPL, AMZ, DELL, GOOG, IBM, INTC, MSFT, SSU
- Time Period: 2013 (or later)-2022
- 17 UN SDGs texts
- Bag-of-words (two-gramms) $\rightarrow$ term-context representations with the pooled vocabulary

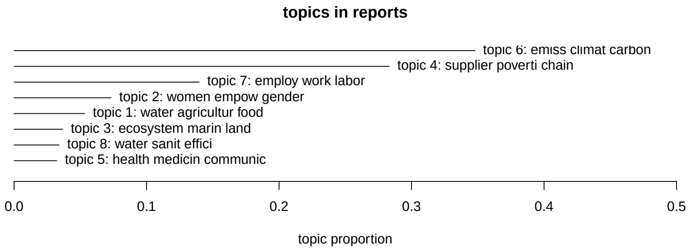
Optimal K and λ

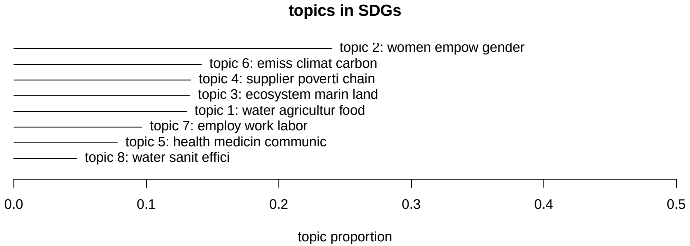
- find the optimal k and λ in a data-driven fashion, via maximizing the **mean coherence**
- mean coherence $\overline{coh}$: the mean of the logratio topic coherence: $\log(\epsilon + TCM_{x,y}) - \log(TCM_{y,y})$ for two terms x, y with TCM being the in-sample term co-occurrence matrix.
- for $K = 8, \lambda = 0$:
  $coh_{sustainability\_reports} = -1.3048, coh_{SDGs} = -7.7671, \overline{coh} = -4.5359$
- for $K = 8, \lambda = 350$:
  $coh_{sustainability\_reports} = -2.6230, coh_{SDGs} = -0.9374, \overline{coh} = -1.7802$

Optimal K and λ

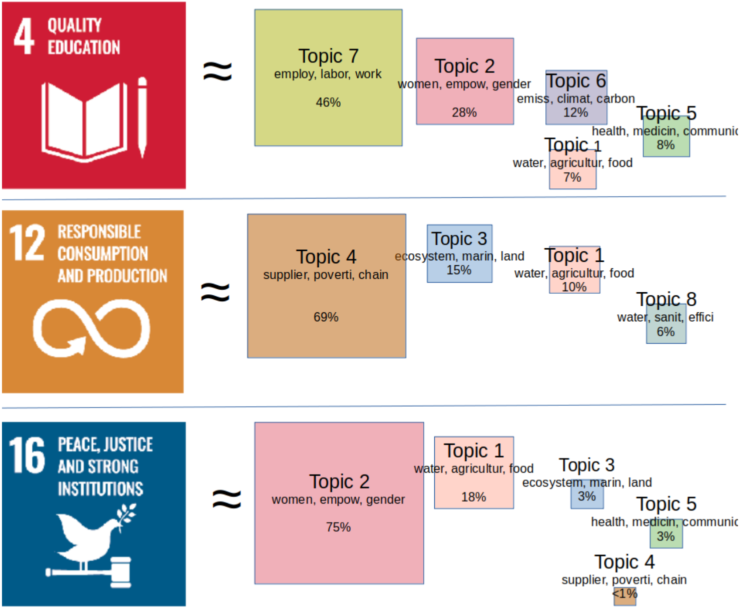


Results: the topics

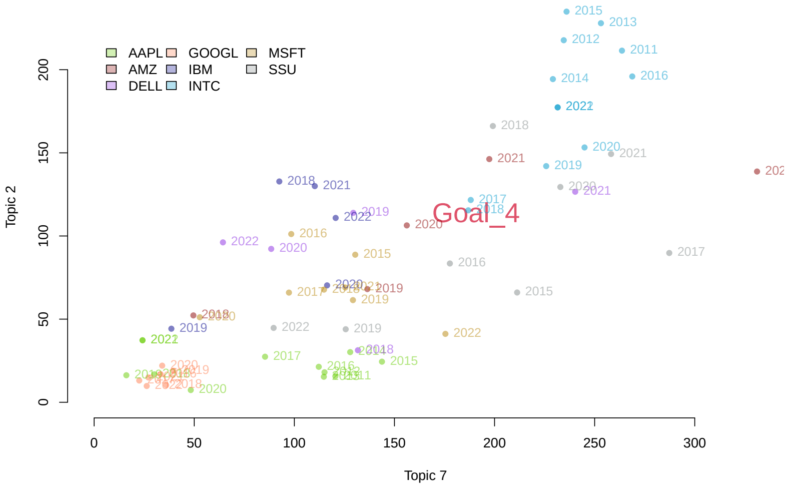




Results: the topics



Results: approximation in two dimensions



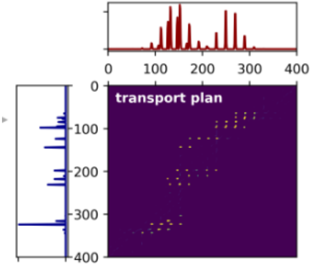
Results: the distributional distances

- Consider reports/SDGs as **distributions**/**histogramms**
- Find an optimal transport plan T^*, such that:

$$\min_T \sum_i \sum_j T_{ij} Cost_{ij},$$

$$s.t. T\mathbf{1}_n = \mathbf{p}, T^\top \mathbf{1}_n = \mathbf{q},$$

where $Cost \in \mathbb{R}^{n \times n}$ is the cost matrix, $T \in \mathbb{R}^{n \times n}$ is the transport plan matrix and $\mathbf{p}, \mathbf{q}$ are (term) probability vectors.



(Source: http://alexhwilliams.info/itsneuronalblog/2020/10/09/optimal-transport/#f5b)

Results: distributional distances

- Consider reports/SDGs as **distributions**/**histogramms**
- Find an optimal transport plan T^*, such that:

$$\min_{T} \sum_{i} \sum_{j} T_{ij} Cost_{ij},$$
$$s.t. T\mathbf{1}_n = \mathbf{p}, T^{\top}\mathbf{1}_n = \mathbf{q},$$

Lee et al. (2022) cosine dissimilarity as cost for optimal transport plan (contextualized mover's distance,CMD):

$$Cost_{ij}^{CMD} = 1 - \cos(x_i, x_j)$$

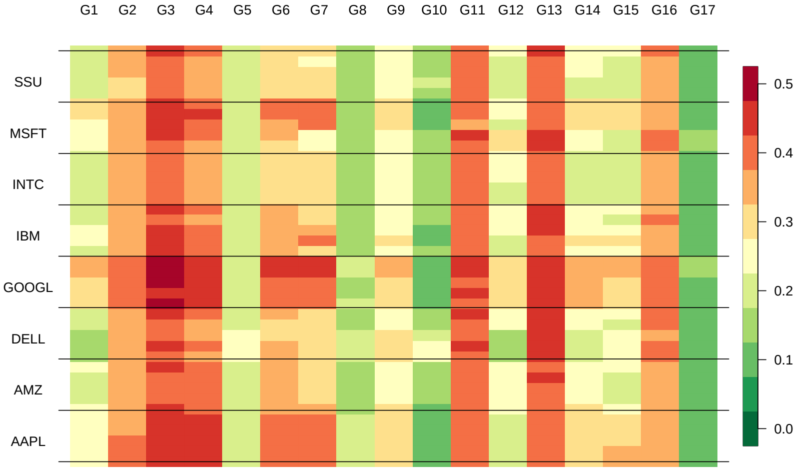
with $x_k, k = 1, \ldots, n$ being the topic-term embedding for the kth term.

- Take the minimized total cost:

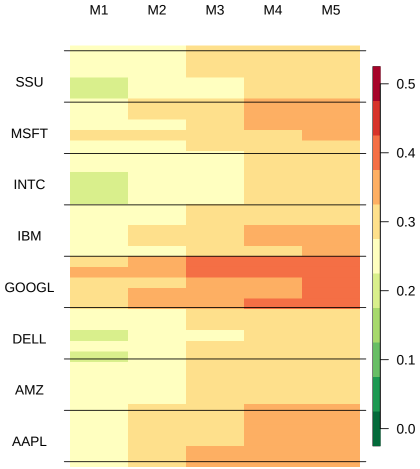
$$Cost^* = \sum_{i} \sum_{j} T^*_{ij} Cost_{ij}^{CMD}$$

to compare reports and SDGs as word distributions.

Results: distributional distances



Results: individual preferences



M1: 1/3 G4 + 1/3 G12 + 1/3 G16

M2: 1/2 G4 + 1/4 G12 + 1/4 G16

M3: 2/3 G4 + 1/6 G12 + 1/6 G16

M4: 3/4 G4 + 1/8 G12 + 1/8 G16

M5: 4/5 G4 + 1/10 G12 + 1/10 G16

Summary

- Topic extraction with side information $\rightarrow$ low dim. representation in a prestructured topic space.
- Projection on a common subspace via non-negative matrix co-factorization, using an algorithm which is easily implemented and delivers interpretable results.
- The resulting topic-term embeddings are used to compare the documents via the optimal transport metric which assists financial decisions under SDGs based preferences.

References I

Berg, Florian, Julian F Kölbel, and Roberto Rigobon. 2022. "Aggregate Confusion: The Divergence of ESG Ratings*." *Review of Finance* 26 (6): 1315–44. https://doi.org/10.1093/rof/rfac033.

Cichocki, Andrzej, Rafal Zdunek, and Shun-ichi Amari. 2007. "Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization." In *Independent Component Analysis and Signal Separation*, edited by Mike E. Davies, Christopher J. James, Samer A. Abdallah, and Mark D. Plumbley, 169–76. Berlin, Heidelberg: Springer Berlin Heidelberg.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2023. "Keyword Assisted Topic Models." https://arxiv.org/abs/2004.05964.

Fang, Yi, and Luo Si. 2011. "Matrix Co-Factorization for Recommendation with Rich Side Information and Implicit Feedback." In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 65–69. HetRec '11. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2039320.2039330.

Fernando, Basura, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. 2013. "Unsupervised Visual Domain Adaptation Using Subspace Alignment." In *2013 IEEE International Conference on Computer Vision*, 2960–67. https://doi.org/10.1109/ICCV.2013.368.

References II

Kang, Hyewon, and Jinho Kim. 2022. "Analyzing and Visualizing Text Information in Corporate Sustainability Reports Using Natural Language Processing Methods." *Applied Sciences* 12 (11). https://doi.org/10.3390/app12115614.

Lee, Seonghyeon, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. "Toward Interpretable Semantic Textual Similarity via Optimal Transport-Based Contrastive Sentence Learning." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5969–79. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.412.

Luo, Ling, Haoran Xie, Yanghui Rao, and Fu Lee Wang. 2019. "Personalized Recommendation by Matrix Co-Factorization with Tags and Time Information." *Expert Systems with Applications* 119: 311–21. https://doi.org/https://doi.org/10.1016/j.eswa.2018.11.003.

Rao, Nikhil, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. 2015. "Collaborative Filtering with Graph Information: Consistency and Scalable Methods." In *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/f4573fc71c731d5c362f0d7860945b88-Paper.pdf.

References III

Watanabe, Kohei, and Yuan Zhou. 2022. "Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches." *Social Science Computer Review* 40 (2): 346–66. https://doi.org/10.1177/0894439320907027.

Zhang, Yupei, Yue Yun, Huan Dai, Jiaqi Cui, and Xuequn Shang. 2020. "Graphs Regularized Robust Matrix Factorization and Its Application on Student Grade Prediction." *Applied Sciences* 10 (5). https://doi.org/10.3390/app10051755.