Directed topic extraction with side information.

Maria Osipenko[1]

14 August, 2023

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

Motivation

- Growing interest to sustainable investments
- Investment decisions based not only on expected return considerations but also relying on individual value system
- Aligning investments with individual preferences - how to quantify sustainability? how to compare investment possibilities?

Motivation

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

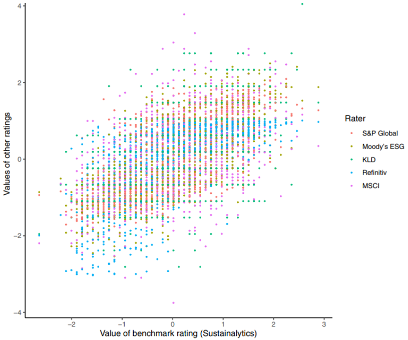


Figure 1: ESG ratings of different providers against a benchmark. Source: Aggregated confusion...

- The weighting systems behind the ratings are partly intransparent and cubersome to understand.

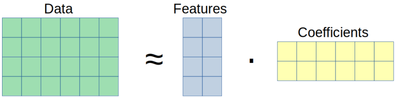
Motivation

- Another source of information easily available to private investors
  - corporate responsibility reports
  - sustainability reports
  - environmental action reports
- A systematic 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 in a low dimensional latent topic space
  - Specific for topic extraction: Latent (probabilistic) Semantic Analysis, Latent Dirichlet allocation (LDA),...
  - General purpose matrix factorization (MF) methods: Principal component analysis, Non-negativ matrix factorization, probabilistic versions and extensions thereof, ...



Methods available

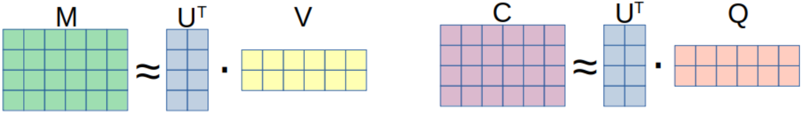
- How to embed known structure or side information in the unsupervised techniques?
  - keyword seeded LDA: Watanabe and Zhou (2022) and Eshima, Imai, and Sasaki (2023)
  - graph regularized MF: Rao et al. (2015) and Zhang et al. (2020) (recommendations)
  - common subspace projection/ subspace alignment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

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

Our approach

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

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



where

- M is the (weighted) term-document matrix for the corporate reports with dimensions $(p \times n)$, where p is the joint vocabulary.
- C is the (weighted) term-document matrix for the sustainability goals with dimensions $(p \times m)$, where p is again the joint vocabulary.
- U is the word-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 topic extraction 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.

Because of the non-negativity of the entries in M and C it makes sense to restrict at least U to be non-negative:

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

Our approach

- why to consider side information? align the topics with a known structure
- why a MCF method? flexible representation in a common low dimensional space
- why Frobenius norm? fast optimization, but other loss specifications are possible.
- why non-negative MCF? enhances the interpretability and sparsity of the resulting topics.

The algorithm

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

Algorithm 1 HALS algorithm for MCF

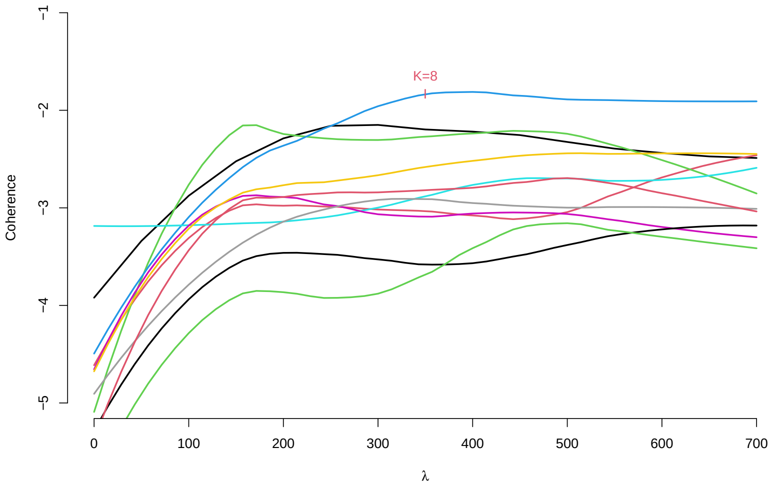
while not converged **do**
  for $k = 1$ to K **do**
    update $V_k \leftarrow \max\left(\frac{U_k(M - U_{-k}^\top V_{-k})}{U_k U_k^\top}, 0\right)$
    update $Q_k \leftarrow \max\left(\frac{U_k(C - U_{-k}^\top Q_{-k})}{U_k U_k^\top}, 0\right)$
    update $U_k^\top \leftarrow \max\left(\frac{(M - U_{-k}^\top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

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

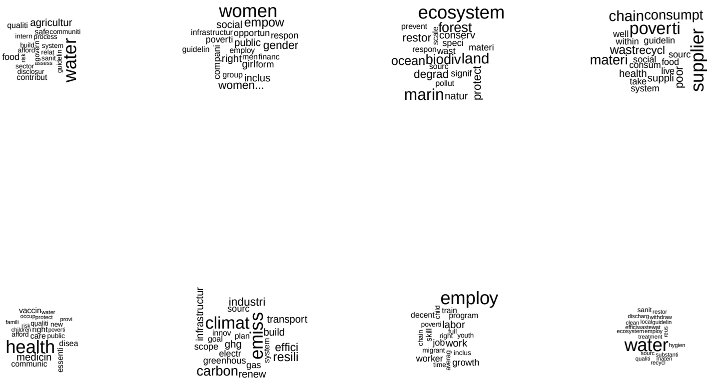
Optimal K and λ

- find the optimal k and λ in a data-driven fashion, via maximizing the average topic coherence
- mean logratio coherence coh_{Corpus} computed as the mean of logratio coherence defined as: $\log(\epsilon + TCM_{x,y}) - \log(TCM_{y,y})$ 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

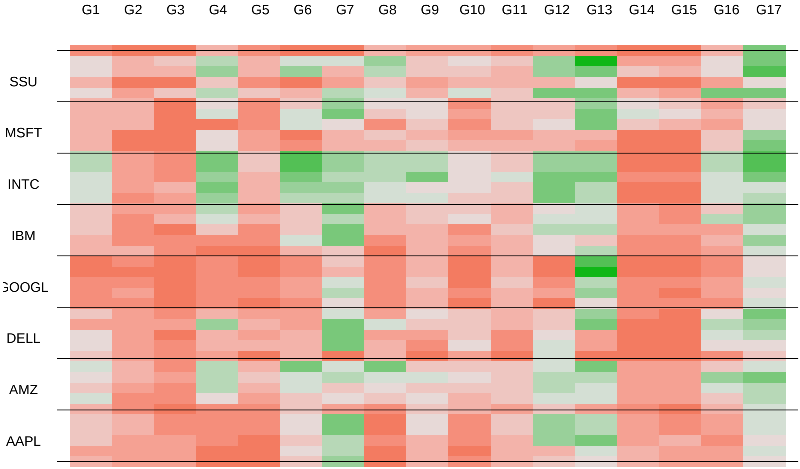


Results: the topics



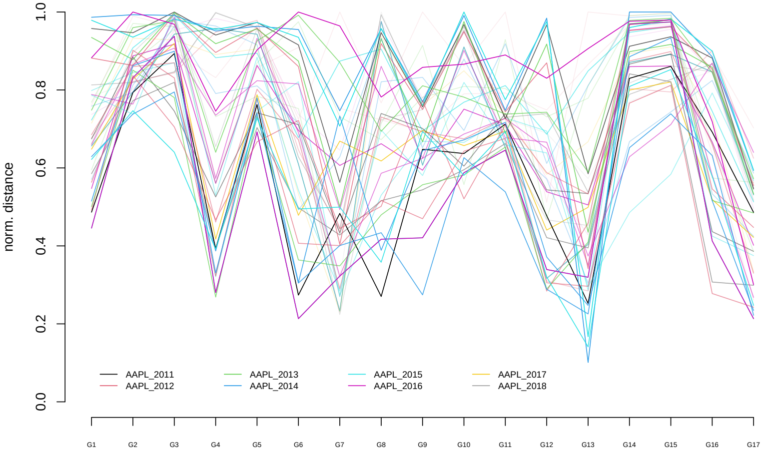
Results: the distances

- distance matrix



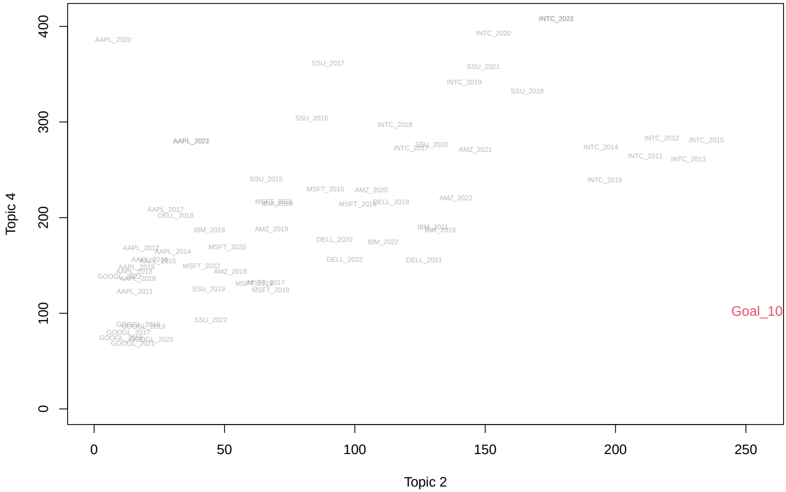
Results: the distances

-parallel coordinate plot



Results: approximation in two dimensions

- plot



Results: individual preferences

- app or pic

Summary

References

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.

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: