Automatic semantic segmentation of the news articles

While much text-mining focuses on short texts on social media, news articles can be lengthy. This raises an issue for the approach we have just described since now-ubiquitous transformer-based models are unable to deal with long sequences, because their self-selection mechanisms scale quadratically with sequence length. *Longformers* (Beltagy et al., 2020) or *Hi-Transformer* (Wu et al., 2021) are meant to tackle this problem, and can accept many more lexical tokens in the input. However, summarizing a long text with a single embedding – which can also be done by averaging the embeddings of successive portions of this text – means losing a lot of relevant fine-grained information. We will thus rely on semantic segmentation to cut each news article into several sections, and conduct content analysis with these sections instead of the whole articles.

To perform segmentation, while recent supervised neural methods (e.g., Lukasik et al., 2020) have shown improved performances, we will choose a robust unsupervised approach based on embedding and computing similarity scores between adjacent blocks to detect topical transitions (Ghinassi, 2021), to take advantage of our multilingual sentence-transformer.

References

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer*. http://arxiv.org/abs/2004.05150

Ghinassi, I. (2021). Unsupervised Text Segmentation via Deep Sentence Encoders: a first step towards a common framework for text-based segmentation, summarization and …. *Proceedings of 2nd International Workshop on Data-Driven Personalisation of Television (DataTV-2021) at the ACM International Conference on Interactive Media Experiences (IMX 2021)*. https://doi.org/10.5281/zenodo.4744399

Lukasik, M., Dadachev, B., Papineni, K., & Simões, G. (2020). Text segmentation by cross segment attention. *Proceedings of the EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing*, *2018*, 4707–4716. https://doi.org/10.18653/v1/2020.emnlp-main.380

Wu, C., Wu, F., Qi, T., & Huang, Y. (2021). Hi-Transformer: Hierarchical Interactive Transformer for Efficient and Effective Long Document Modeling. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, *2*, 848–853. https://doi.org/10.18653/v1/2021.acl-short.107