

Democratising data science on corpora: automated knowledge extraction and visualisation at ease

Evelina Di Corso

evelina.dicorso@polito.it

Department of Control and Computer Engineering,
Politecnico di Torino, Torino, Italy

Tania Cerquitelli

tania.cerquitelli@polito.it

Department of Control and Computer Engineering,
Politecnico di Torino, Torino, Italy

ABSTRACT

Nowadays, large volumes of textual data are continuously collected at an ever-increasing rate in various modern applications, ranging from social networks (e.g. Twitter, Facebook) to digital libraries (e.g. Wikipedia). We are in an age of data-intensive science and we are witnessing the unprecedented generation of large corpora.

The analysis of these collections is **challenging**, as it is a multi-step process in which the data scientists tackle the complex task of configuring the analytics system to transform data into **actionable knowledge** to effectively support the decision-making process. A plethora of algorithms are currently available for performing a given data analysis phase, but for each one the specific parameters have to be manually set, and the obtained results validated by a domain expert. Moreover, real datasets are also characterised by an inherent sparseness and variable distributions, and their complexity increases with the data volume. Thus, a proper combination of different analytics algorithms should be defined to correctly model data under analysis. These activities are very **time-consuming** and require **a lot of expertise** to achieve the best trade-off between the quality of the result and the execution time.

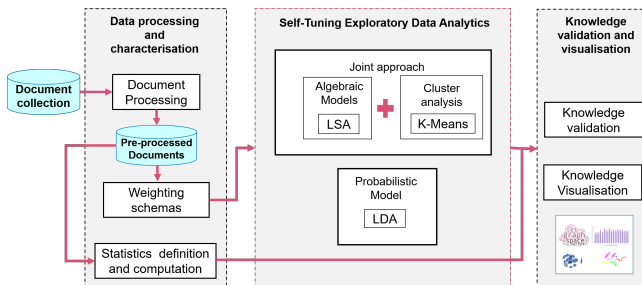


Figure 1: The ESCAPE System Architecture

To streamline the complexity of the data analytics pipeline on textual corpora and allowing all people (including those that never study data science algorithms) to benefit from data-driven methodologies, we have designed and developed an **automated data analytics engine** [3] to effectively and efficiently analyse large collections of textual data with minimal user intervention. The new engine, named **ESCAPE** (Enhanced Self-tuning Characterisation of document collections After Parameter Evaluation, see Figure 1) includes two different solutions to address document clustering

[4, 6] and topic modelling [7]. In each proposed solution, ad-hoc **self-tuning strategies** have been integrated to automatically configure the specific algorithm parameters, as well as the inclusion of **novel visualisation techniques** and **quality metrics** [1, 2] to analyse the performances of the methodologies and help people to understand and to easily interpret the discovered knowledge. We believe the proposed visualisation (Fig.2) approaches along with the automated data analytics engine could democratise the effective exploitation of data science in analysing textual corpora [5].

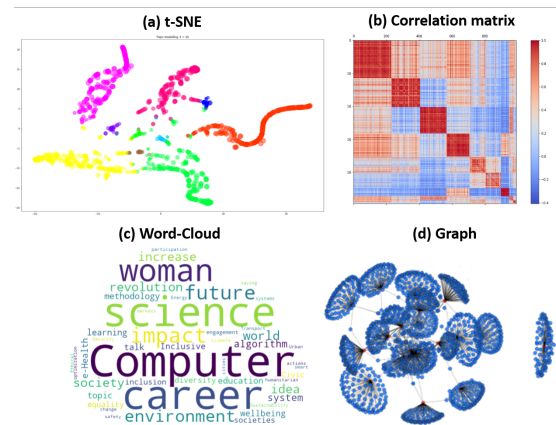


Figure 2: Knowledge visualisation examples

Possible **future extensions** concern the integration of other (i) methodologies, and (ii) visualisation techniques.

REFERENCES

- [1] Tania Cerquitelli, Evelina Di Corso, Francesco Ventura, and Silvia Chiusano. 2017. Prompting the data transformation activities for cluster analysis on collections of documents. In *Proceedings 25th Italian Symposium on Advanced Database Systems*.
- [2] Tania Cerquitelli, Evelina Di Corso, Francesco Ventura, and Silvia Chiusano. 2017. Data miners' little helper: data transformation activity cues for cluster analysis on document collections. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*. ACM, 27.
- [3] Evelina Di Corso. June 2019. Text miner's little helper: Scalable self-tuning methodologies for knowledge exploration. (June 2019).
- [4] Evelina Di Corso, Tania Cerquitelli, and Francesco Ventura. 2017. Self-tuning techniques for large scale cluster analysis on textual data collections. In *Proceedings of the Symposium on Applied Computing*. ACM, 771–776.
- [5] Evelina Di Corso, Stefano Proto, Tania Cerquitelli, and Silvia Chiusano. 2019. Towards automated visualisation of scientific literature. In *European Conference on Advances in Databases and Information Systems*. Springer - In press.
- [6] Evelina Di Corso, Francesco Ventura, and Tania Cerquitelli. 2017. All in a twitter: Self-tuning strategies for a deeper understanding of a crisis tweet collection. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 3722–3726.
- [7] Stefano Proto, Evelina Di Corso, Francesco Ventura, and Tania Cerquitelli. 2018. Useful ToPIC: Self-Tuning Strategies to Enhance Latent Dirichlet Allocation. In *2018 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 33–40.