On the Interplay Between Business Process Management and Internet-of-Things — A Systematic Literature Review

Francesca De Luzi, Francesco Leotta, Andrea Marrella, Massimo Mecella

Business & Information Systems Engineering (2024)

Appendix (available online via http://link.springer.com)

Appendix

The Latent Dirichlet allocation (LDA) technique by Blei et al. (2003) is one of the most effective ones for latent topic distribution within a corpus Chauhan and Shah (2021). LDA is a generative probabilistic model of a corpus where the documents are represented as a random mixture of latent topics, each characterized by a mix of words. This appendix further describes how the LDA technique, as reported in Sect. 4.2, was used to identify the topics in the 93 studies selected by the SLR.

The topic identification process consists of a preliminary pre-processing and topic modeling. As a primary step in data processing, it was necessary to analyze the studies to recover the relevant information such as title, abstract, conclusions, and a summary of the study. Then, a plain text file was created, composed of 93 rows, each for each study. Once the four parts of each study were selected, a pre-processing phase of the texts was conducted, to clean and standardize them and to adequately train the machine learning model.

The first operation is tokenization, in which each text is cleaned by separators and special characters (e.g. "=") and then broken down into sentences. Finally, each token is subject to lowercasing and punctuation removal operations. At this point, we can proceed with removing the so-called stopwords. In human language, some terms have no semantic meaning but only serve to intercalates between periods, such as articles, conjunctions, prepositions, and all those words of non-semantic contribution. Therefore, these are usually removed from the text during the pre-processing phase to decrease the number of words and increase the quality, emphasizing words that carry a more significant amount of information.

Before starting the training phase, it is necessary to specify the number of topics the model will have to identify within the various texts. We ran a grid search[11] to understand the best value of several topics and compared the clarity of the multiple models trained with different values of several topics. The measure used to assess the clarity of a model is the coherence score, a measure of the distance between the various clusters. The graph in Fig. 5 highlights that the model with four topics performs best. The model's accuracy drops dramatically when the number of topics is set to 5-6 and goes back when the

[11] Grid search, stands for GridSearch Cross Validation and is a complete search to check all the specified value combinations.

number of topics exceeds 7. To interpret the topics, we used a web-based interactive visualization that helps analysts quickly inspect the topic-term relationships of an LDA model, the LDAV*is*. The interactive version of our model built with LDAV*is* is shown in Fig. 6. The four topics are represented as circles in a 2D plane, where the circle distances indicate the similarity or difference of the designated topics. Lastly, the topic's importance, which depends on the number of studies associated with the topic, is represented by the areas of the circles.

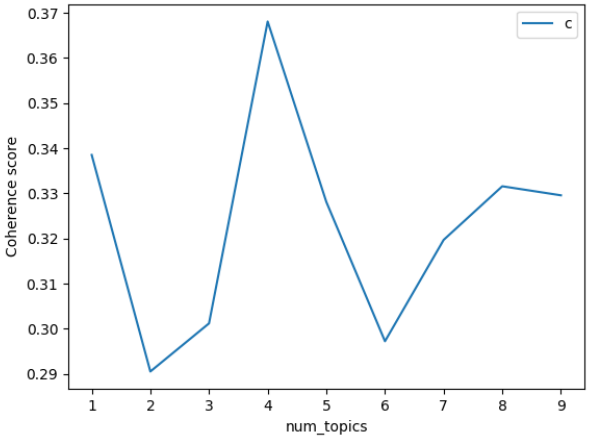

Figure 5: The purity of the model on varying the number of topics to isolate.

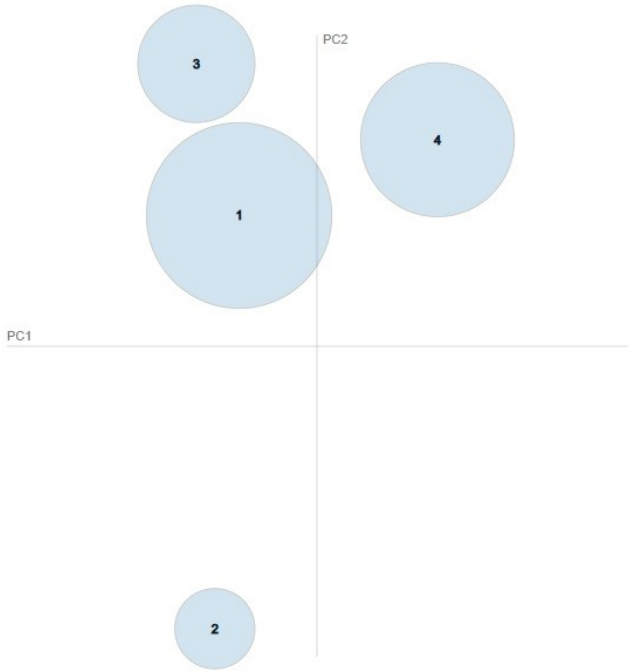

Figure 6: Inter-topic distance map.