Improving a taxonomy through Natural Language Processing

submitted by Radio-Canada

Radio-Canada publishes between 450 and 600 new contents every day. These contents are created and categorized by our edimasters and journalists in every part of the country. The current taxonomy is used in order to place (categorize) the relevant content on our various digital platforms: it is not used to describe the nature of the content. Our goal is to develop a better understanding of the topics within our content and of the interests of our audience (in order, for instance, to design advanced research tools and algorithms for content recommendation): thus we wish to enrich our taxonomy while limiting the work load of our teams.

Our objective is to improve our taxonomy: (1) through Natural Language Processing techniques, in order to extract the essence of a text and obtain a representation of the entities contained therein; (2) through the creation of “logical groups” of similar entities.

We own a catalogue of more than 17,000 textual contents that have been categorized; they can be used by the team members. We can also share several artifacts (research, strategy, models) with the team to support the work of the participants.