Document Classification API

Via document classification, a document is automatically assigned to a category, out of a large set of predefined categories. For example, the document can be about "Sport", or "Science and Technology" or "Politics", etc. A document can actually belong to multiple categories, with higher or lower affinity.

Technically, the task of document classification is carried out by a *machine learning model*, trained on a *large corpus* of example documents, from a predefined *categorization scheme*. The definitions of categories are inherently *domain-specific*, as it is hard to define a scheme that encompasses "all themes" that text documents can be about. We opted for a categorization that best suits news data (also suitable to blogs, twitter, etc.). The technical documentation presents the approach in detail.