Topic modeling: Through the Eyes of our Working Example

OUR STORY

Feature engineering is the process of using domain knowledge related to your data to select and identify *features* in the data that can be leveraged by machine learning. That is not a hard definition, because sometimes it is used in a context where features are transformed for machine learning and domain knowledge is included to create new features.

Unfortunately, engineered features are not easy to identify for large data sets. When there are many features, generally only a small number play an important role when it comes to prediction. Furthermore, when creating new features, domain insight is even more difficult to fold into the model when there are hundreds or thousands of features to keep track of.

Natural language processing (NLP) is an area of machine learning that makes heavy use of feature engineering. NLP is a problem domain that is not trivial, nor is it overly complicated by having too many features to reasonably work with. One process for identifying features in text data is known as *topic modeling*. Topic modeling is an unsupervised process that enables us to identify clusters of features that exist in text data. For this case study, we will use topic modeling to generate insights from unstructured text.

This is a task that data scientists are often assigned. The result of these types of experiments will be shared with domain experts during playbacks to further engineer features that are relevant when it comes to particular business opportunities. Accordingly there will be a focus on visualizing the findings in anticipation of communication with domain stakeholders.

THE DESIGN THINKING PROCESS

Feature engineering is yet another key skill you will be employing (along with dimensionality reduction and bias mitigation) as a data scientist during the *Ideate*, *Prototype*, and *Testing* phases of a design thinking project. Feature engineering in relation to text analysis is especially important because most large enterprises are looking for ways to deal with the vast amounts of unstructured data (text) available in the organization.