research@eulytix.eu

CoFE Datathon 2022

Challenge 2: Proposal Clustering

Methodological brief

Summary

The essence of topic modeling is to try to identify topics based on the semantic similarity of the texts. A number of models are used in the literature and in practice, the most common of which are perhaps Latent Diritchlet Allocation (LDA)[1] and Correlated Topic Models (CTM).[2] These methods are considered to be somewhat outdated by some, their primary disadvantage is that the analyst has to determine the number of potential topics in advance. This is exacerbated by the lack of a universally accepted method for determining the number of topics. Therefore, we opt to employ the simple, intuitive and elegant technique called "Top2Vec" introduced by Dimo Angelov.[3]

Top2Vec in a nutshell

The Top2Vec method first generates vectors representing words and documents. These result in a semantic space that hosts words and documents, which can be hundreds of dimensions. In more technical terms Top2Vec starts out by producing a joint embedding of words and proposals. Several methods exist for this embedding task. We opt to use Doc2Vec,[4] as it is quite well-known and language agnostic.

The resulting semantic space sparse and is of high dimensions. To make the detection of clusters more feasible, the next step in Top2Vec is to map this space into a lower dimensional one. Angelov's solution uses *Uniform Manifold Approximation and Projection for Dimension Reduction* (*UMAP*)[5] for this task.

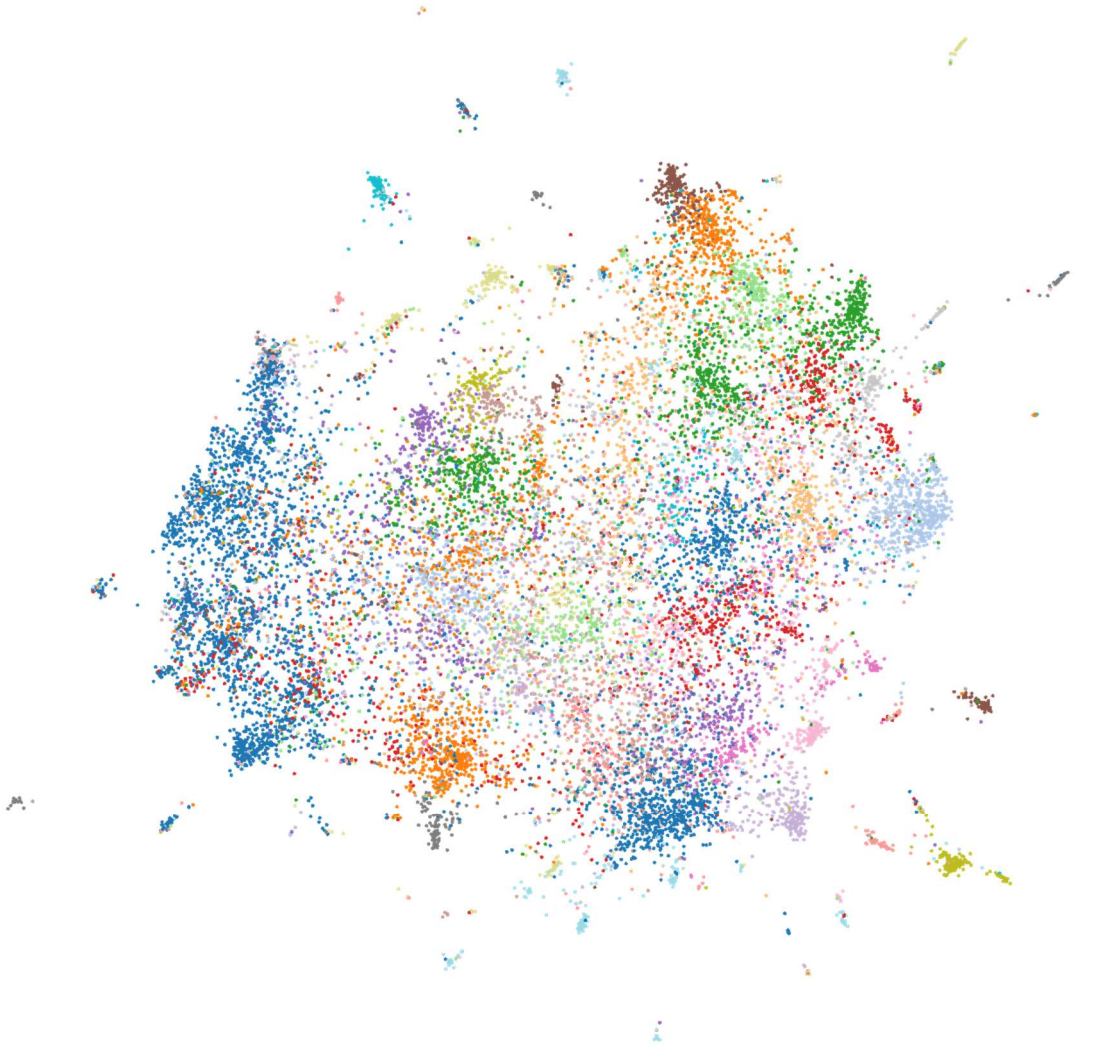
Now that we have reduced the dimensionality of the semantic space, we can identify clusters with the use of HDBSCAN.[6] The identified clusters may be aggregated to obtain the desired number of clusters (50,20 and 10 in this case). An illustration of the resulting clusters is presented in Figure 1.

---

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." J. Mach. Learn. Res. 3(0): 993–1022.
[2] Lafferty, John, and David Blei. 2006. "Correlated Topic Models." In Advances in Neural Information Processing Systems, eds. Y. Weiss, B. Schölkopf, and J. Platt. MIT Press.
[3] Angelov, Dimo. 2020. "Top2Vec: Distributed Representations of Topics." https://arXiv.org/abs/2008.09470.
[4] Le, Quoc V., and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." https://arxiv.org/abs/1405.4053.
[5] McInnes, Leland, John Healy, and James Melville. 2020. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv:1802.03426
[6] McInnes, L., J. Healy, and S. , Astels. 2017. "Hdbscan: Hierarchical Density Based Clustering." The Journal of Open Source Software 2.

research@eulytix.eu

Figure 1: Illustration of the UMAP-reduced document vectors of the CoFE dataset and the identified clusters. Different colors indicate different clusters. Note that different clusters may have the same colour.



Notes:

Though this method is language agnostic, we decided to perform this exercise on the English versions of the proposals. The reasoning behind this is that this method would most likely identify the different languages as clusters.

With this method, no removal of stopwords, stemming/lemmatizing is necessary. The reason is that stopwords are present in every cluster, hence they are identified as outliers that play no role in the determination of cluster centroids. Furthermore, the use distributional representations of words and documents makes stemming/lemmatization unnecessary.