文本挖掘科研初探

BERTopic: Neural topic modeling with a class-based TF-IDF procedure

张建章

阿里巴巴商学院
杭州师范大学

2023-02-22

目录

BERTopic 是一种基于文本聚类的主题挖掘技术：① 利用基于 Transformer 的 embeddings 技术将文档向量化为 embeddings；② 使用层次密度聚类 (HDBSCAN) 对文档向量进行聚类；③ 使用类簇 TF-IDF 挑选可以表示类簇主题的词汇。

论文: Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint: 2203.05794 (2022). [pdf]

代码: BERTopic-github 主页.

该算法本质上是组合式创新，其基本思路是通过聚类文本嵌入向量，来简化文本主题挖掘，同时得到可解释的主题词汇 [1,2]。

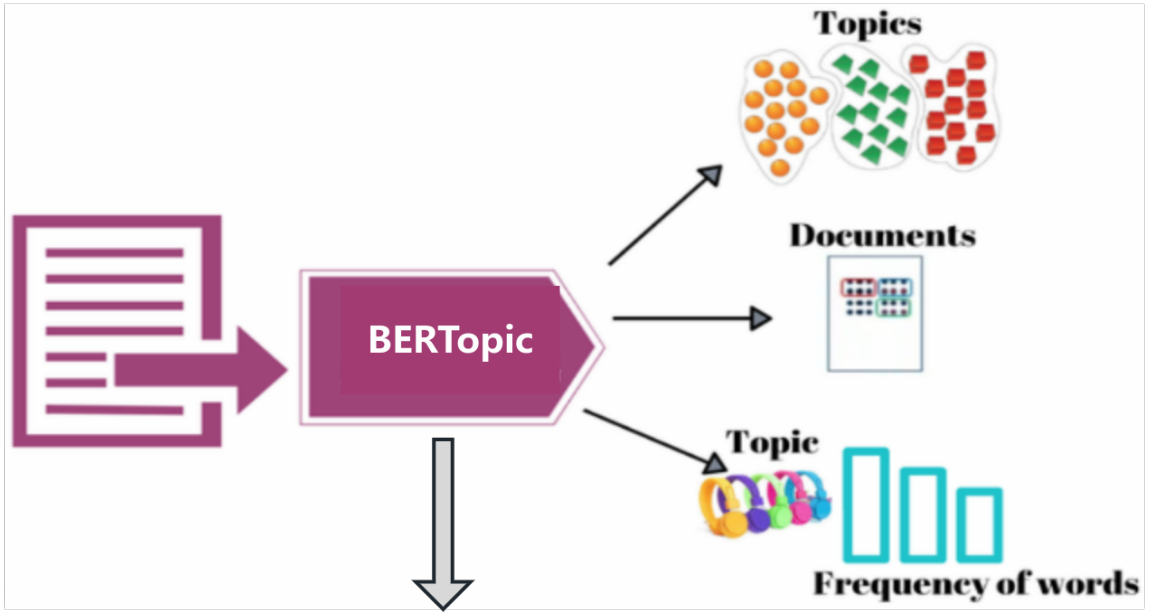
该思路不基于 pLSA、LDA 等传统的概率生成主题模型，但由于其结果亦包含主题划分和主题词语，故可采用已有的主题一致性 (topic coherence) 和主题词多样性 (topic diversity) 衡量指标进行结果的定量评估与对比，可使用论文 [3] 提出的评测框架进行横向 (comparison) 和消融 (ablation) 评测。

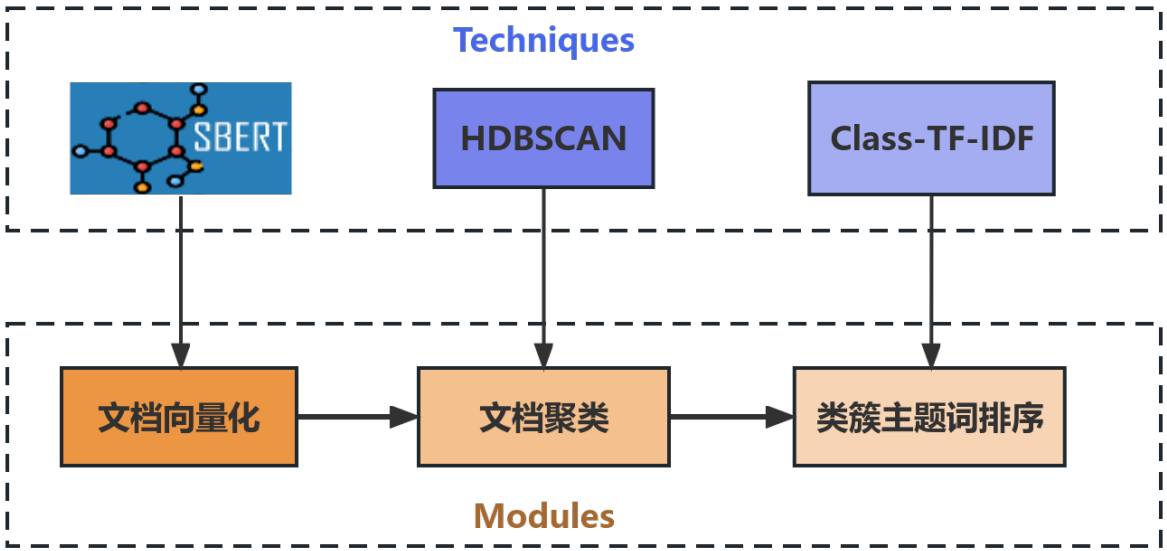
[1] Sia, Suzanna, Ayush Dalmia, and Sabrina J. Mielke. "*Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!*." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1728-1736. 2020. [pdf]

[2] Angelov, Dimo. "*Top2vec: Distributed representations of topics*." arXiv preprint arXiv:2008.09470 (2020). [pdf]

[3] Terragni, Silvia, et al. "*Octis: comparing and optimizing topic models is simple!*." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2021. [pdf]

BERTopic 方法框架





模块 1-文档向量化

借助预训练语言模型 (pretrained language models) 将文档进行向量化表示，可以使用：① 第一代词向量，如，word2vector，fasttext，GloVe；② 第二代词向量，Transformer based language models，如 BERT，RoBERT 等；③ 以及文档向量预训练模型，如 Doc2Vector，sentence transformer (SBERT) 等。

BERTopic 使用SentenceTransformers获取整个文档的嵌入向量表示。该框架整合了句子 (sentence)、文档 (text)、图像 (image) 嵌入向量的 SOTA 模型，包含多语种模型，其中的模型可用于语义匹配、自然语言问答、图像搜索等多种应用场景。其原始论文如下：

论文: Reimers, N. and Gurevych, I., 2019, November. "*Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*". In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982-3992). [pdf]

代码: SBERT-github 主页

模块 2-文档聚类

使用 HDBSCAN (Hierarchical Density-Based Spatial Clustering) 对文档向量进行聚类。相比于原始的 DBSCAN 算法，HDBSCAN 可以找到具有不同密度的类簇。

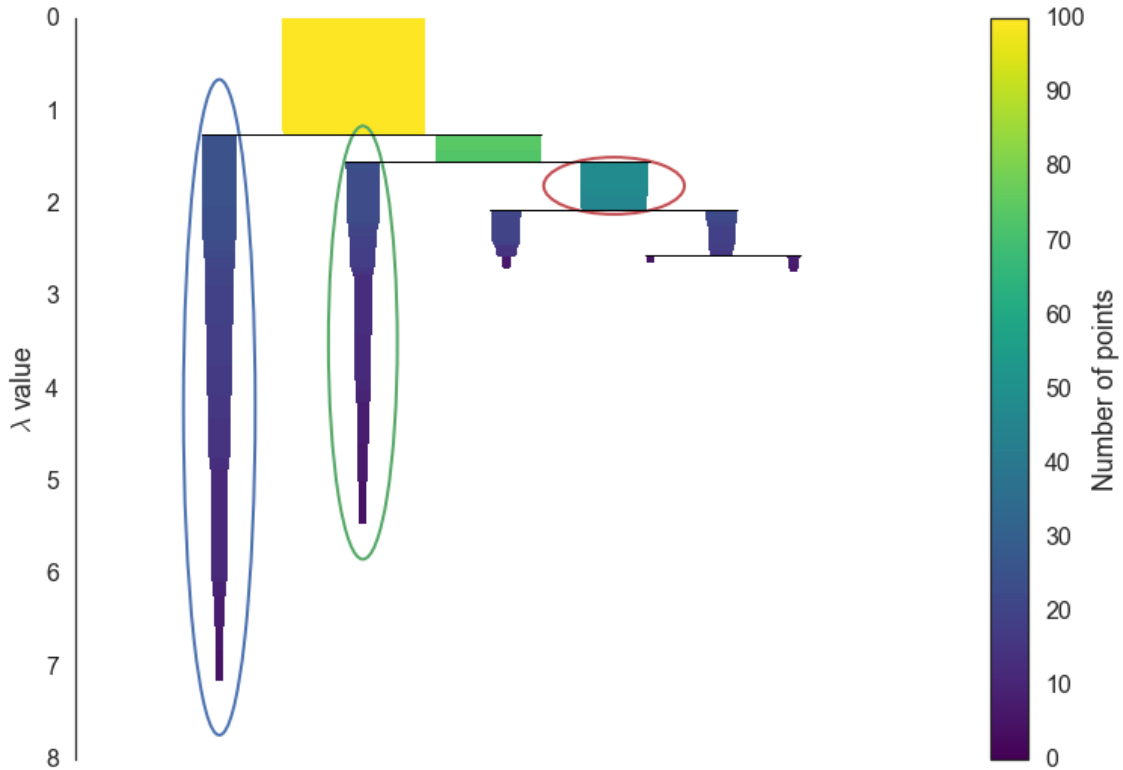


图 1: hdbscan 示意图

为聚类结果中的每个类簇分配一个主题。

基本思想：对不同的 ϵ 执行 DBSCAN 并对结果进行整合，以找到可提供最佳 ϵ 稳定性的聚类，ϵ 是 DBSCAN 聚类算法中确定两个点是否为邻居的距离阈值。

参考论文：

R. Campello, D. Moulavi, and J. Sander, “Density-Based Clustering Based on Hierarchical Density Estimates” In: Advances in Knowledge Discovery and Data Mining, Springer, pp 160-172. 2013. [pdf]

McInnes L, Healy J. “Accelerated Hierarchical Density Based Clustering” In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, pp 33-42. 2017. [pdf]

McInnes, L., Healy, J., & Astels, S. (2017). “hdbscan: Hierarchical density based clustering”. Journal of Open Source Software, 2(11), 205. [pdf]

代码：HDBSCAN-github 主页

聚类使用的输入是文档嵌入向量的降维表示，使用了 UMAP 向量降维算法，以在低维投影空间中保留更多的高维数据中的全局和局部特征。

模块 3-类簇主题词排序

计算每个类簇中词语的 TF-IDF，具体使用类别 TF-IDF (c-TF-IDF)，计算公式如下：

$$w_{t,c} = tf_{t,c} \times \log(\frac{A}{tf_t})$$

其中，$tf_{t,c}$ 表示词语 w 在类簇 c 中的词频，A 表示所有的类簇数量，tf_t 表示包含词语 w 的类簇数量。上式第二部分本质是逆类别频率。实际计算中，将每个类簇中的文档拼接 (concatenate) 成一个文档。

对每个类簇中的词语按照 c-TF-IDF 进行排序，取出 Top k 个词作为该类簇的主题词。

Datasets: 20 NewsGroups; BBC News; Trump Tweets (当选前和当选后，具有时间维度的数据集)，前两个数据集均为文本挖掘领域常用的基准 (Benchmark) 数据集。

Baselines: LDA, NFM (非负矩阵分解), Topic2Vector, CTM (Contextual Topic Modeling)，前两个为传统的基于概率和线性代数的主题模型，第三个为与 BERTopic 同类的聚类主题挖掘模型，第四个为将嵌入向量与传统主题模型相结合 (incorporation) 的方法。

Measures: *TC* 和 *TD* 分别表示主题一致性 (Topic Coherence) 和主题多样性 (Topic Diversity)。

Table 1: 横向对比结果

	20 NewsGroups		**BBC News**		**Trump**	
	TC	TD	TC	TD	TC	TD
LDA	.058	.749	.014	.577	-.011	.502
NMF	.089	.663	.012	.549	.009	.379
T2V-*MPNET*	.068	.718	-.027	.540	-.213	.698
T2V-*Doc2Vec*	.192	.823	.171	.792	-.169	.658
CTM	.096	.886	.094	.819	.009	.855
BERTopic-*MPNET*	.166	.851	.167	.794	.066	.663

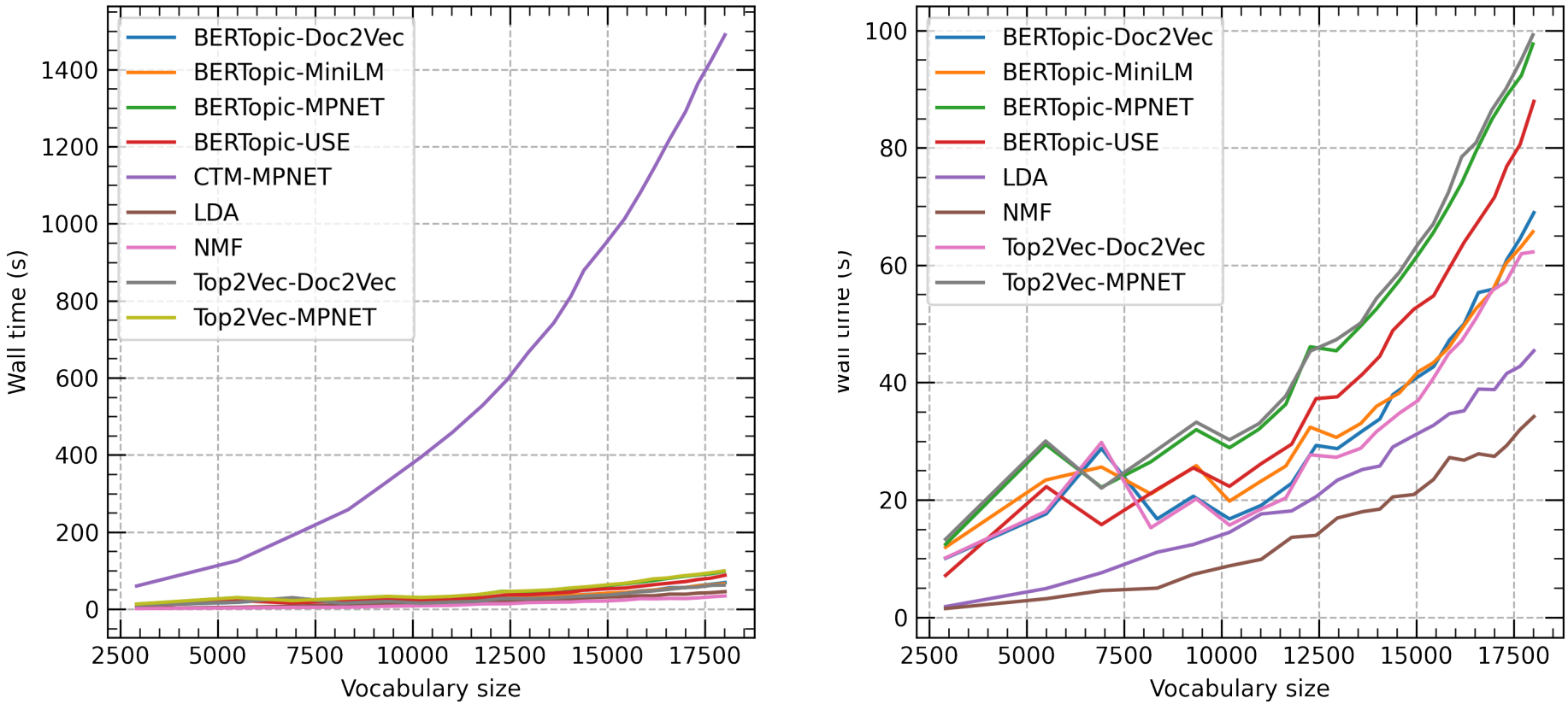


Figure 1: Computation time (wall time) in seconds of each topic model on the Trump dataset. Increasing sizes of vocabularies were regulated through selection of documents ranging from 1000 documents until 43000 documents with steps of 2000. **Left**: computational results with CTM. **Right**: computational results without CTM as it inflates the y-axis making differentiation between other topic models difficult to visualize.

综合考虑性能和时间 (trade-off between performance and running time)，可选择 MiniLM 作为模块 1 中的文本嵌入向量模型。

```
# 模型训练
from bertopic import BERTopic
from sklearn.datasets import fetch_20newsgroups

docs = fetch_20newsgroups(subset='all',  remove=('headers',
↪  'footers', 'quotes'))['data']

topic_model = BERTopic()
topics, probs = topic_model.fit_transform(docs)

# 输出类簇
topic_model.get_topic_info()

# 输出类簇0的主题词
topic_model.get_topic(0)

# 主题结果可视化
topic_model.visualize_topics()
```

更多用法请参考 BERTopic 文档，并搭配 `help` 函数学习使用。

论文: 以会议论文为主，建议参考 CCF 推荐国际学术会议和期刊目录中的人工智能目录，代表性会议有 ACL、EMNLP 等；

代码：Github，Towards Data Science，Kaggle 等；

实验: 推荐使用在线代码平台 Kaggle，本地做实验推荐使用 linux 系统或 Mac OS 系统。

1. 在理解 BERTopic 框架的基础上，使用 App 隐私政策作为输入文档进行主题挖掘，如主题挖掘结果不符合实际需求，则对该框架中的模块进行适应性改进，如，替换 Embedding 模型，替换聚类方法等；

2. 将 BERTopic 方法用简洁的算法伪代码表示，伪代码格式可参考 HAC 算法流程伪代码；

3. 使用大 O 法计算 BERTopic 模型的时间复杂度。

THE END