Big Data and Automated Content Analysis (6EC)
Week 6: »Unsupervised machine learning«
Monday

Anne Kroon
a.c.kroon@uva.nl, @annekroon

May 11, 2023

UvA RM Communication Science

Before we start: Are there questions?

Today

Unsupervised machine learning

- Should one still use LDA?
- State-of-the-art approaches to topic modelling

Final project

Unsupervised machine learning

Today: Unsupervised machine learning for text

Using topic models

You got your model – what now?

1. Assign topic scores to documents
2. Label topics
3. Merge topics, throw away boilerplate topics and similar (manually, or aided by cluster analysis)
4. Compare topics between, e.g., outlets
5. or do some time-series analysis.

Example: Tsur et al., 2015

Unsupervised machine learning

Should one still use LDA?

The popularity of LDA

In the last decade, LDA has become *extremely* popular in the social sciences due to

- easy-to-use R and Python packages
- its promise to not require (a) manual (qual or quant) analysis; (b) annotations for SML; (c) creation of dictionaries etc.
- a bit of a “cool new technique” image

The popularity of LDA

But there is no silver bullet!

Unfortunately,

- validating topic models is hard – and many (most) studies don't do it (well);
- there are so many choices and parameters, in combination with no simple and definite evaluation metric, that it is very hard to justify why a particular model is chosen;
- experience shows that it often "doesn't work" ⇒ it's quite normal to have many uninterpretable or ambigous topics;
- The smaller the dataset, the less likely it is to work
- LDA tends to also pick up pecularities that don't matter and outliers

Solutions?

There are some extensions on classical LDA, in particular:

- Author-topic models
- Structural topic models (STM) (Roberts et al., 2014)
- Dynamic topic models (Blei & Lafferty, 2006)

These allow covariates (e.g., add info on who wrote a text) to improve the model, or allow to account for the changing use of words and topics over time.

Also, there are techniques for validation available (e.g., topic intrusion and/or word intrusion tasks).

Solutions?

But some we can't solve everything.

- It's still BOW.
- We cannot incorporate any language knowledge from larger, pre-trained datasets (e.g., via embeddings)

⇒ If we think of the performance leap that we observe with Transformers in other areas, we have all reason to assume that we can do better.

Unsupervised machine learning

State-of-the-art approaches to topic modelling

Let’s bring in embeddings and Transformers!

Using embeddings and transformers for topic modelling

For example:

- top2vec (Angelov, 2020), which embeds *topic vectors* in the same space as document vectors and word vectors
- Contextualized Topic models (Bianchi, Terragni, & Hovy, 2021; Bianchi, Terragni, Hovy, et al., 2021), with a lot of code examples at https://contextualized-topic-models.readthedocs.io/en/latest/introduction.html
- . . .

Using embeddings and transformers for topic modelling

For example:

- top2vec (Angelov, 2020), which embeds *topic vectors* in the same space as document vectors and word vectors
- Contextualized Topic models (Bianchi, Terragni, & Hovy, 2021; Bianchi, Terragni, Hovy, et al., 2021), with a lot of code examples at https://contextualized-topic-models.readthedocs.io/en/latest/introduction.html
- . . .

Using embeddings and transformers for topic modelling

For example:

- top2vec (Angelov, 2020), which embeds *topic vectors* in the same space as document vectors and word vectors
- Contextualized Topic models (Bianchi, Terragni, & Hovy, 2021; Bianchi, Terragni, Hovy, et al., 2021), with a lot of code examples at https://contextualized-topic-models.readthedocs.io/en/latest/introduction.html
- . . .

BERTopic (Grootendorst, 2022)

“In this paper, we introduce BERTopic, a topic model that leverages clustering techniques and a class-based variation of TF-IDF to generate coherent topic representations. More specifically, we first create document embeddings using a pretrained language model to obtain document-level information. Second, we first reduce the dimensionality of document embeddings before creating semantically similar clusters of documents that each represent a distinct topic. Third, to overcome the centroid-based perspective, we develop a classbased version of TF-IDF to extract the topic representation from each topic. These three independent steps allow for a flexible topic model that can be used in a variety of use-cases, such as dynamic topic modeling.”

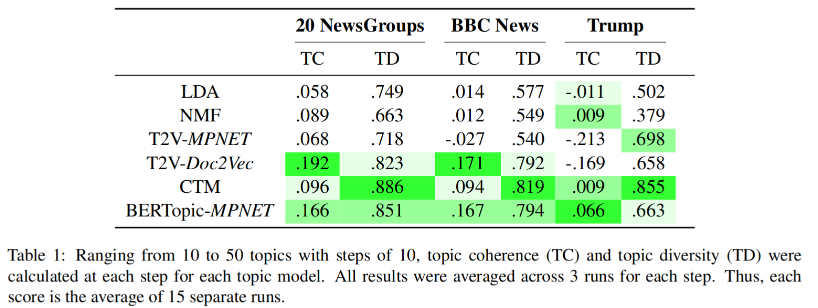
(for details, read the paper)

BERTopic (Grootendorst, 2022)

Let's look at specific examples, for instance: https://maartengr.github.io/BERTopic/getting_started/quickstart/quickstart.html

but also the visualization capabilites:
https://maartengr.github.io/BERTopic/getting_started/visualization/visualization.html#visualize-topics-per-class

Much more coherent topics than LDA!



	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

Table 1: Ranging from 10 to 50 topics with steps of 10, topic coherence (TC) and topic diversity (TD) were calculated at each step for each topic model. All results were averaged across 3 runs for each step. Thus, each score is the average of 15 separate runs.

(And no need to set k! And there is a dedicated “outlier topic” called -1!)

Are there downsides?

Of course!

- By definiton, much more “black-box”-y than BOW approaches
- Risk of biases introduced by LLM
- Much more resource-hungry (you probably want to do this with a GPU (e.g., on CoLab)

To conclude: LDA is an interesting starting point – but if I were to start an unsupervised topic analysis model now, I'd go for BERTopic.

Final project

Exercising with unsupervised machine learning for text

Two notebooks in this week's folder: LDA and BERTopic.

But in particular, look at the BERTopic website for more examples!

It’s time to think about your final projects!

- Let’s look at the course manual!
- Talk to me about your ideas!
- Main point: It needs to be sth you like, and it needs to cover techniques the course!

References

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 759–766. https://doi.org/10.18653/v1/2021.acl-short.96

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1676–1683. https://www.aclweb.org/anthology/2021.eacl-main.143

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, 113–120.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064–1082.

Tsur, O., Calacci, D., & Lazer, D. (2015). A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1629–1638.