Improving topic models with segmental structure of texts

Nikolay Skachkov, MSU (nikolaj-skachkov@ya.ru)
Konstantin Vorontsov, MIPT (vokov@forecsys.ru)

01/06/2018

Topic Modeling

Given: A document collection. Each document is a bag-of-words:

$$p(w|d) = \frac{n_{wd}}{n_d}$$

Find: Documents and words per topic distribution (topical embeddings):

$$\phi_{wt} = p(w|t), \quad \theta_{td} = p(t|d)$$

- Topic modeling task can be interpreted as a soft clusterization of documents and words. Solution of this task in non-unique.
- ARTM[1] allows to impose additional problem-specific criteria on the topic model.

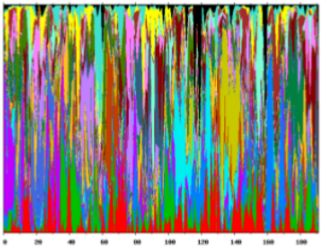
[1] *BigARTM — open-source project for topic modeling (http://bigartm.org/)*

Topic Segmentation

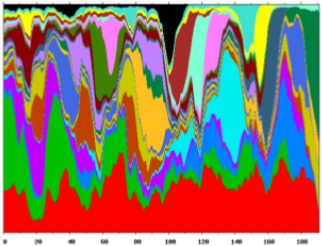
Assumption:

- Each document is represented as a sequence if segments;
- Words within a segment share the same small set of topics.

Topical representation of a document (x-axis: sentence number in text; y-axis: topic distribution)



standard topic model (LDA)



topic model with segment boundaries estimation (TSM[2])

[2] *L. Du, W. Buntine, Topic Segmentation with a Structured Topic Model (2013)*

Objective

Our aim is to model intra-document topic behavior with a good quality and high sparsity.

Applications:

1. Big document analysis;
2. Small collections analysis;
3. Intra-document navigation;
4. Topical retrieval;
5. Document summarization;
6. Dialog analysis;
7. Topical scenarios building.

Topic Tiling

TopicTiling[3]– segmentation algorithm, that uses any topic model to estimate segment boundaries in the texts.

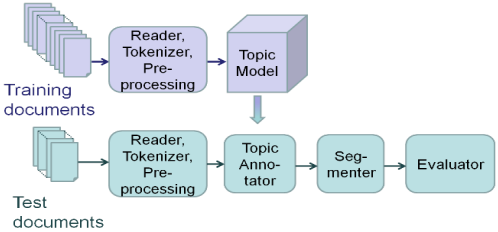
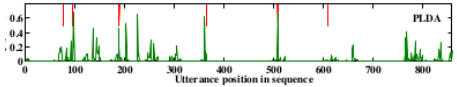


Figure 1: Basic concept of text segmentation using Topic Models



Segment boundary probabilities for each sentence ending

[3] *M. Riedl, C. Biemann* Text Segmentation with Topic Models. 2012.

ArtPostScience and segmentation quality.

To evaluate segmentation quality of the methods an artificial collection was built. In artificial documents golden standard segment boundaries are provided.

ArtPostScience[4]:

- Artificial documents are built by concatenating full source documents from PostScience collection;
- Sequential segments in artificial documents share different topics.

P_k and **WindowDiff** metrics are used to evaluate models. Both of them compare a given segmentation of a document with golden standard. WindowDiff is more sensitive to small segments. A value close to 0 denotes a perfect segmentation quality for both metrics.

[4] ArtPostScience: https://yadi.sk/d/fSswtwqV3SbsCD

Models description

1. ARTM + sparse Θ — model that reduces the number of topics in documents topical embeddings. It is built in a bag-of-words assumption.
2. ARTM + SentenceSparse — model that reduces the number of topics in each sentence.
3. ARTM + SegmentSparse — model that reduces the number of topics in each segment and merges sequential segments if they share same topics.

Models evaluation

Results of segmentation quality on test documents:

Model	WindowDiff metrics	P_k metrics
TopicTiling	0.258	0.145
ARTM + sparse Θ	0.173	0.100
ARTM + SentenceSparse	0.159	0.099
ARTM + SegmentSparse	**0.155**	**0.095**

Evaluation of sparsity of documents topical embeddings:

Model	Θ sparsity
ARTM + sparse Θ	96%
ARTM + SegmentSparse	98.5%

As we can see ARTM + SegmentSparse provides the best results in segmentation using less number of topics.

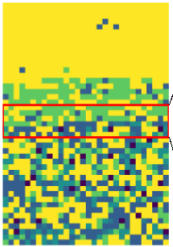
Document topical segmentation example

Topic 1
галактика
система
слово
объект
буква

Topic 2
китайский
культурный
возраст
система
учёный

Topic 3
закон
государство
язык
словарь
слово

SegmentSparse Model:



Поэтому неудивительно, что в таком регионе возникает двуединый этнос — общность хуася, сочетающая традиции аустрических народов и сино-тибетцев. — (segment boundary – **detected**) — 4 февраля 1722 года Петром I была принята Табель о рангах — закон, регламентировавший порядок государственной службы в Российской Империи и определявший соответствие гражданских, военных и придворных чинов.

PLSA Model:

Поэтому неудивительно, что в таком регионе возникает двуединый этнос — общность хуася, сочетающая традиции аустрических народов и сино-тибетцев. — (segment boundary – **not detected**) — 4 февраля 1722 года Петром I была принята Табель о рангах — закон, регламентировавший порядок государственной службы в Российской империи и определявший соответствие гражданских, военных и придворных чинов.

Document topical segmentation on real texts

Example of a document from original PostScience processed by SegmentSparse Model:

... Казанская губерния, наоборот, вошла по просьбе Коржинского. Это интересная историческая ситуация, но постепенно, уже к 20-м годам XX века средняя Россия охватывает территорию от Ярославской и Костромской губерний на севере до Воронежской и Саратовской на юге. Вот эта вся территория находится в пределах европейской части России на левобережье Волги. В среднем, по некоторым подсчётам, природная флора этого региона насчитывает примерно 4,5 тысячи видов, очень немного. Для сравнения флора Турции, которая меньше по площади, включает больше 15 тысяч видов. Связана бедность флоры, с одной стороны, с тем, что это равнинная территория...

SegmentSparse model found two different segments. The first topical segment contains historical information when the second segment is more about natural features of the region. So this segmentation of the document is justified.

Conclusion

- A new TopicSegmentation regularizer was added to BigARTM. That means that this regularizer can be used with another assumptions about topic model;
- Proposed method increases segmentation quality;
- Proposed method increases sparsity of documents topical embeddings;
- Proposed method doesn't complicate the structure of model training;
- **This method allows to build informative topical embeddings within documents.**