Extracting Reliable Topics Using Topic Model Ensembles

Alex J. Loosley[1], Alexandre D. Salles[1], Stephan Sahm[1], and Juan Bernabé-Moreno[2]

[1]Data Reply GmbH, Munich, Germany; [2]University of Granada, Department of Computer Science and Artificial Intelligence, Granada, Spain

Introduction and Motivation

- As part of several recent industry projects, my colleagues and I have used topic models to extract topics from text.
- A recurring theme in these projects is the question, which topics extracted from a topic model are reliable?
- Unreliable topics can be attributed to:

  Too little structure in the data to find a suitable compression

  Learning convergence issues (e.g. not enough EM steps for LDA, not enough NMF gradient descent iterations for LSI)

  The local minimum an algorithm finds if optimization converges (Fig. 1)
- In this work, we begin to address these challenges by training a topic model ensemble and identifying reliable topics as those that reproducibly occur in the ensemble.
- Topics in the ensemble that form compact clusters over words are said to reproducibly occur.

Random State 1 | Random State 2

Figure 1. Two 20-topic topic models trained on customer service text from an online forum show varied results over different random initializations (visualizations generated using pyLDAvis)

Algorithm

Reliable topics are extracted as follows:

1. Concatenate topic term distributions from each model in the ensemble:

$$\boldsymbol{\mathcal{T}} = \begin{bmatrix} \boldsymbol{\mathcal{T}}_1 \\ \dots \\ \boldsymbol{\mathcal{T}}_R \end{bmatrix} = \begin{bmatrix} \mathcal{T}_{1,1} \\ \mathcal{T}_{2,1} \\ \dots \\ \mathcal{T}_{K,1} \\ \dots \\ \mathcal{T}_{1,2} \\ \mathcal{T}_{2,2} \\ \dots \\ \mathcal{T}_{K,R} \end{bmatrix} \tag{1}$$

where $\mathcal{T}_{k,r} = P_{W|T,\mathcal{M}}(w|t_k, m_r)$ is the topic term distribution of the k^{th} topic of the r^{th} topic model.

2. Calculate pairwise directionally masked cosine distances between each row of the concatinated topic term distributions (Eqn. 1) to form the asymmetric distance matrix:

$$M_{ij} = \delta(\mathcal{T}_i, \mathcal{T}_j) = 1 - \text{mask}(\mathcal{T}_i; \mathcal{T}_i) \cdot \text{mask}(\mathcal{T}_j; \mathcal{T}_i) \tag{2}$$

where *i* and *j* are topic indices, and mask($\mathcal{T}_i$; $\mathcal{T}_j$) is a function that:

1. applies a mask to $\mathcal{T}_i$ corresponding to keeping the largest 98% of $\mathcal{T}_j$'s terms by weight
2. renormalizes the masked copy of $\mathcal{T}_i$

$M_{ij=0}$ if distribution $\mathcal{T}_i$ is contained within distribution $\mathcal{T}_j$, up to a scalar normalization constant.

3. Apply a variant DBSCAN with an extra check-back step to find clusters based on the asymmetric distance matrix M_{ij}. The check-back step dictates that a newly proposed core only assumes the label of its parent if at least 75% of the return distances between the new core and all cores with the parent's label are less than the DBSCAN distance scale.

Experiments

Synthetic Data:

- Synthetic topic term distributions, $P_{W|T}$ were generated as shown below:

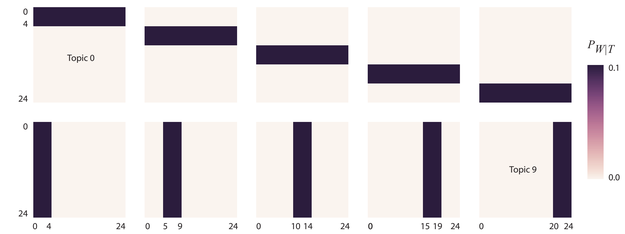


Figure 2. Synthetic text topic term distributions. Each pixel on the heatmap represents one of 25^2 unique terms, indexed on the x and y axes from **0** to **24**. Each topic consists of **100** works forming the corresponding horizontal and vertical bars shown here.

- 250 homogeneous (single-topic) 100-term synthetic texts were generated for each topic by randomly sampling 100 terms from the corresponding topic term distribution
- A topic ensemble consisting of 25 10-topic topic models (250 topics learned altogether) was trained.
- The reliable topics learned recapitulated the synthetic data (Fig. 3).

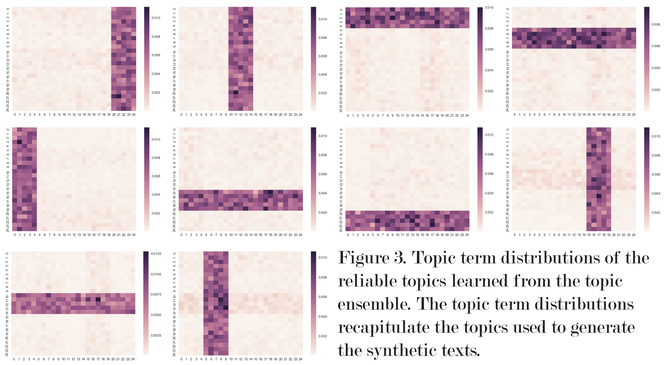


Figure 3. Topic term distributions of the reliable topics learned from the topic ensemble. The topic term distributions recapitulate the topics used to generate the synthetic texts.

- Other experiments show that the method can be robust to choice of number of topics. Hence, the topic ensemble has capacity to learn the number of topics (under conditions still being investigated).

Natural Language (German online forum text):

- Scraped online customer support forum posts from a large company
- Non-German texts were removed (langdetect), and the remaining texts cleaned via POS filtering (scrapy.de), stopword removal, and synonym mapping
- 20x 50-topic LDA topic models were trained with different random states
- Topics from the topic ensemble where clustered using the call-back variant of DBSCAN (see Algorithm panel), to identify reliable topics (Fig. 4)

All Topics in Ensemble | Stable Topics Only (outliers removed)

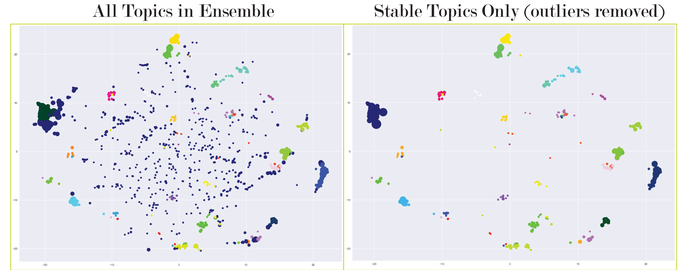
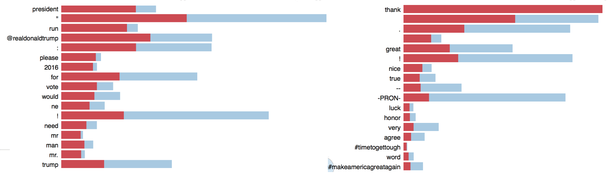


Figure 4. tSNE mappings of the concatentated topic term distributions from the topic ensemble. The ensemble consisted of 20x 50-topic topic models (1000 topics learned altogether). Each point represents a topic, point size represents the marginal topic likelihood on the training data, and point colour represents the cluster label. The left plot shows all 1000 topics. The right plot shows the same topics filtered to keep only the cluster cores. These latter points represent the reliable topics.

Natural Language (English tweets):

- Scraped 31,000 Donald Trump tweets
- Tweets cleaned by encoding with tf-idf
- Only two reliable topics learnable (experiments with ›10 models, ›10 topics per model)



- It remains to be validated whether 31,000 short Trump tweets truly contain only two reliable topics.

Summary and Work in Progress

- This work proposes a topic ensemble and clustering technique to robustly identify reliable topics that summarize a set of texts.
- Topic model ensembles are found to robustly find all topics generating synthetic text and we are currently experimenting to find the conditions under which the topic ensemble works on natural language.

  The topic ensemble shows capacity to learn every topic from *k*-topic synthetic texts, even when than number of topics in the ensemble is not equal to *k*.
- Convergence as the number of models increases must be investigated and bagged topic ensembles need to be tested.

Acknowledgements

- Data Reply GmbH., for supporting this work and generously funding my industry researcher position at the Machine Learning Summer School (Tübingen, 2017)
- Max Planck Institute for Intelligent Systems for providing me an opportunity to engage with the academic machine learning community
- Please communicate with Alex Loosley about this work, a.loosley@reply.de