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, or artefacts, can be attributed to:
  - Too little sample data about a particular topic
  - Convergence issues (EM for LDA, NMF for LSI)
  - The random initial contitions from which topic models are trained (Fig. 1)
- These challenges are addressed 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 the topic term distributions from each model in the ensemble to form the concatenated topic term distribution matrix (CTTD)

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

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 the set of pairwise directionally masked cosine distances (DMCD) between each row of the CTTD to form the following asymmetric distance matrix:

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

where i and j are topic model indices, and mask$(\mathcal{T}_i; \mathcal{T}_j)$ is a mask function that masks $\mathcal{T}_i$ based on $\mathcal{T}_j$. Specifically, mask$(\mathcal{T}_i; \mathcal{T}_j)$ sorts $\mathcal{T}_j$ values from largest to smallest and sets mask indices corresponding to the first 97.5% of the sorted distribution weight. The corresponding mask is applied to $\mathcal{T}_i$ and both distributions are renormalized. $M_{ij=0}$ if distribution j is contained within distribution i, up to a scaler multiplier.

3. A variant of DBSCAN called check back DBSCAN (cbDBSCAN) was applied to find clusters from the asymmetric distance matrix $M_{i,j}$. The algorithm for cbDBSCAN is the same as DBSCAN except there is a check-back step once a core has been identified. Specifically, a core is labeled the same as its parent if and only if at least 75% of the return distances between the core and its parent and parent neighbours are less than ε, the DBSCAN density parameter.

Experiments

Synthetic Data:

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

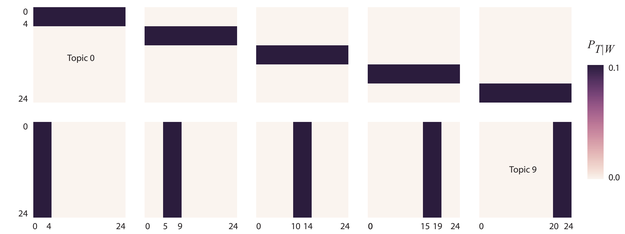


Figure 2. Synthetic text topic term distributions. Each pixel on the heatmap represents one of 25^2 unique terms, indexed on the x and y axes from 0 to 24. Each topic consists of 100 works forming the corresponding horizontal and vertical bars shown here.

- 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).

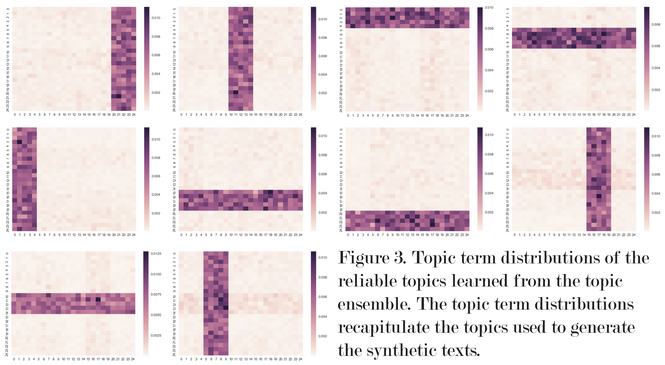


Figure 3. Topic term distributions of the reliable topics learned from the topic ensemble. The topic term distributions recapitulate the topics used to generate the synthetic texts.

- Other experiments not shown here 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 we are still determining.

Natural Language (customer support forum sized German texts):

- Scraped online customer support forum from a large telecommunications company
- Non-German texts were removed (langdetect), with 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 in topic ensemble where clustered using cbDBSCAN (see Algorithm panel), to identify stable topics (figure below)

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

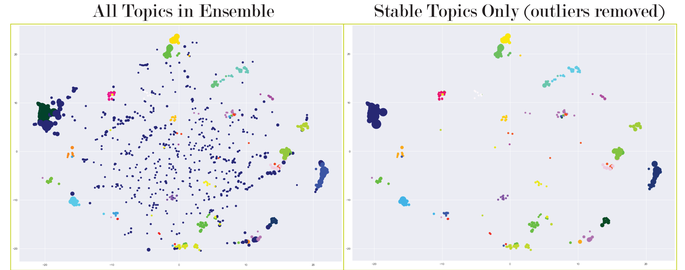
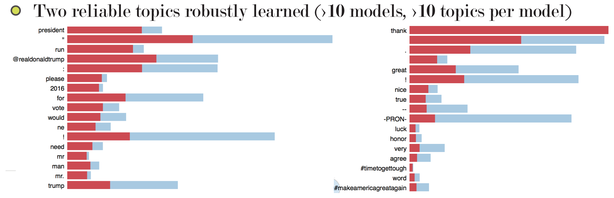


Figure 4. tSNE mappings of the concatentated topic term distributions from the topic ensemble. The ensemble consisted of 20, 50-topic topic models (1000 topics learned altogether). Each point represents a topic, point size represents the marginal topic likelihood on the training data, and point colour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

Natural Language (tweet sized English texts):

- Scraped 31,000 Donald Trump tweets
- Tweets cleaned simply by encoding with tf-idf
- Two reliable topics robustly learned (>10 models, >10 topics per model)



- Some other variations in text preprocessing also resulted in a "make America great again" reliable topic. It remains to be validated whether 31,000 short Trump tweets truly contain only two or three 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 mapping the conditions in which this technique works on natural language.
  - Trained on K-topic synthetic data, the topic ensemble shows capacity to learn all K-topics, even when than number of topics for each topic model is less than K.
- Convergence as the number of models increases must be investigated, and bagged topic ensembles need to be tested.

Acknowledgements

- Data Reply GmbH., for generously funding my tenure at the Machine Learning Summer School (Tübingen, 2017)
- Max Planck Institute for Intelligent Systems and Machine Learning Summer School, for helping me live out my dream as Machine Learning contributor to the community
- Please communicate with Alex Loosley about this work, a.loosley@reply.de.