Motivation

Ensemble LDA

For unsupervised text exploration, many use LDA over other topic models because the learned latent representation is often more interpretable compared to representations from other models. When we use such models in industry projects, a reoccurring question is, “which topics extracted from the topic model are reliable?” In this context, reliable means, if an *k*-topic model were to be trained again and again on the same data, which topics would repeatably be found in the topic model ensemble? The reliable topics typically represent word distributes from which the bag of words was actually generated, whereas the other topics are typically noise or artifacts of a model that needs to fit *k* topic-term distributions to the data, regardless of what *k* is.

Unreliable topics, or artefacts, can be attributed to the following:

1. Too little sample data about a particular topic
2. Algorithmic convergence problems (EM for LDA, NMF for LSI)

   In our experiments, a commonly reoccurring event is for one topic-term distribution to converge to the union of two separate reliable topic-term distributions. We call such a topic a **composite topic**.
3. The random initial condition from which the topic models are trained (see Fig. 1)

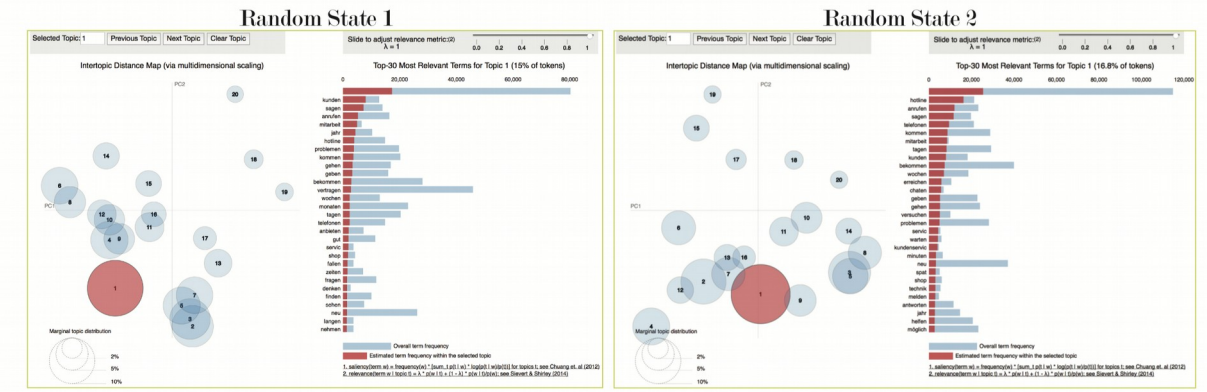

Figure 1. Two 20-topic topic models trained on customer service text from an online forum show varied results over different random initializations (visualizations generated using pyLDAvis)

EnsembleLDA, provided in this PR, proposes a solution to the unreliable topic problem.

1. Train an ensemble of topic models:

As the name indicates, the algorithm trains an ensemble of LDA topic models.

2. Cluster the topics together:

Topics are clustered together using a special version of DBSCAN that we call check-back DBSCAN (CBDBSCAN). CBDBSCAN is a variant designed specifically to robustly cluster LDA topic distributions together. The algorithm for CBDBSCAN is the same as DBSCAN except we use an asymmetric distance matrix and there is a special check-back step once a core has been identified, in order to determine if the core should have the same cluster label as its parent. In short, the asymmetric checkback allows the clustering mechanism to ignore unreliable **composite topics** that occur when training LDA models (see *eLDA_algo_overview.pdf* for more details)..

3. Choose to keep only reliable topics

Only topics with enough cores from the CBDBSCAN are labeled as reliable (see Fig. 2)

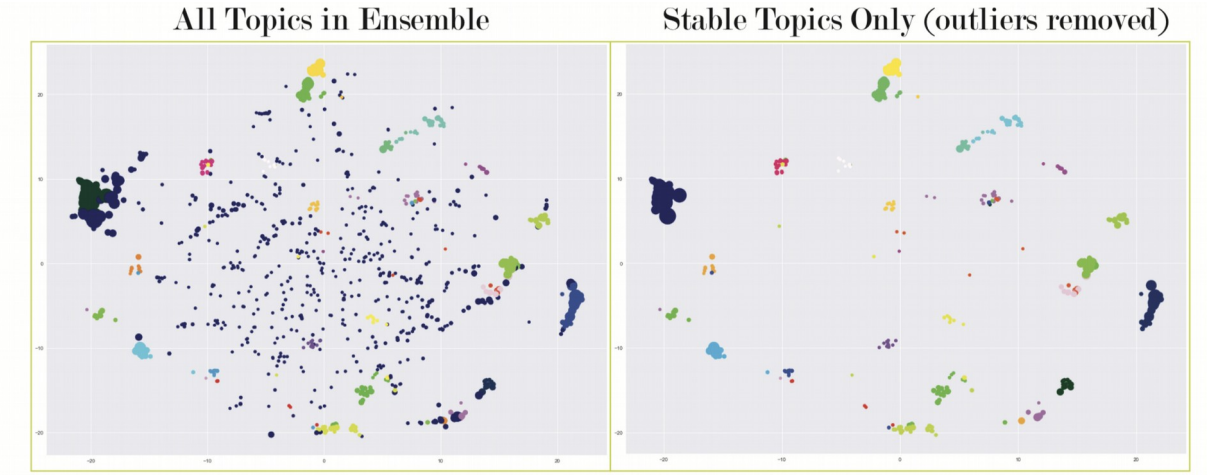

Figure 2. All topic-term distributions from each topic model in the topic ensemble trained on customer support forum text (embedded into 2D with tSNE). The ensemble consists of 20, 50-topic topic models (1000 topics learned altogether). Each point represents a topic, point size represents the marginal topic likelihood on the training data, and point colour represents the cluster label. The left plot shows all the topics. The right plot shows the same topics filtered to keep only the labeled-cores found by CBDBSCAN. These average topic-term distributions for each respective topic cluster represent the reliable topics.

4. Insert reliable topics in a gensim LdaModel object so the user can use the full power of gensim to infer the topics of new documents

Our bachelor thesis paper shows the robustness of this model to choices of hyper-parameters. One of the breakthroughs besides finding the reliable topics is that the model does so robustly against various choices of *k*, the number of topics. The rule of thumb is, the large the model space in the ensemble (i.e. the more compute power one throws at EnsembleLDA), the more

likely the result will be accurate. We have shown this result on synthetic text and through a range of experiments on real text.