Stability of topic modeling via matrix factorization

Mark Belford*, Brian Mac Namee, Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland

ABSTRACT

Topic models can provide us with an insight into the underlying latent structure of a large corpus of documents. A range of methods have been proposed in the literature, including probabilistic topic models and techniques based on matrix factorization. However, in both cases, standard implementations rely on stochastic elements in their initialization phase, which can potentially lead to different results being generated on the same corpus when using the same parameter values. This corresponds to the concept of "instability" which has previously been studied in the context of k-means clustering. In many applications of topic modeling, this problem of instability is not considered and topic models are treated as being definitive, even though the results may change considerably if the initialization process is altered. In this paper we demonstrate the inherent instability of popular topic modeling approaches, using a number of new measures to assess stability. To address this issue in the context of matrix factorization for topic modeling, we propose the use of ensemble learning strategies. Based on experiments performed on annotated text corpora, we show that a K-Fold ensemble strategy, combining both ensembles and structured initialization, can significantly reduce instability, while simultaneously yielding more accurate topic models.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Topic models aim to discover the latent semantic structure or topics within a corpus of documents, which can be derived from co-occurrences of words across the documents. Popular approaches for topic modeling have involved the application of probabilistic algorithms such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). More recently, Non-negative Matrix Factorization approaches (Lee & Seung, 1999) have also been successfully applied to identify topics in unstructured text (Arora, Ge, & Moitra, 2012; Kuang, Choo, & Park, 2015).

The standard formulations of both the LDA and NMF algorithms include stochastic elements in their initialization phase, prior to an optimization phase which produces a local solution. This random component can affect the final composition of the topics found and the rankings of the terms that describe those topics. This is problematic when seeking to capture a definitive topic modeling solution for a given corpus and represents a fundamental *instability* in these algorithms – different runs of the same algorithm on the same data can produce different outcomes. This problem has been widely studied in the context of partitional clustering algorithms

such as k-means, which tends to converge to one of numerous local minima, depending on the choice of starting condition (Bradley & Fayyad, 1998). It has long been recognized as a significant drawback of such algorithms and a substantial number of works exist which attempt to address the issue (*e.g.* Kuncheva & Vetrov, 2006; Pena, Lozano, & Larranaga, 1999).

In the case of topic modeling, instability can manifest itself in two distinct aspects. The first can be observed when examining the topic descriptors (*i.e.* the top terms representing each topic) over multiple runs. The term rankings may change considerably, where certain terms may appear or disappear completely between runs. Secondly, issues of instability can also be observed when examining the degree to which documents have been associated with topics across different runs of the same algorithm on the same corpus. In both cases, such inconsistencies can potentially alter our interpretation and perception of a given topic model. Also, it is clear that any individual run should not be treated as a "definitive" summary of the underlying topics present in the data.

Generally speaking, in the comparative evaluation of topic modeling approaches, researchers tend to focus on either the coherence of the topic descriptors (Newman, Lau, Grieser, & Baldwin, 2010) or the extent to which the topics accurately coincide with a set of ground truth categories or human annotations (Kuang et al., 2015). However, few researchers have considered the evaluation of different approaches from the point of view of their stability across multiple runs.

* Corresponding author.
E-mail addresses: mark.belford@insight-centre.org (M. Belford), brian.macnamee@ucd.ie (B. Mac Namee), derek.greene@ucd.ie (D. Greene).

In this paper we quantitatively assess the extent to which standard randomly-initialized NMF and LDA algorithms are unstable with respect to the topics that they produce on a diverse collection of text corpora. To do this we propose measures that capture the two distinct aspects of instability outlined above. We then focus on addressing the issue in the context of matrix factorization, exploring the use of strategies that involve improved initialization and ensemble learning. In particular, we propose a new combined approach, motivated by the traditional concept of *k*-fold cross-validation, which can yield stable results while also often producing more accurate and coherent models.[1]

The rest of the paper is structured as follows. In Section 2 we provide an overview of relevant work in topic modeling and the more general area of cluster analysis. In Section 3 we discuss the problem of topic model instability in more detail, describing three new measures to quantify instability in topic models. In Section 4 we propose ensemble approaches to address the issue, which are subsequently evaluated on ten different text corpora in Section 5. Finally in Section 6 we conclude the paper with ideas for future work.

2. Related work

2.1. Topic modeling

Topic models attempt to discover the hidden thematic structure within an unstructured collection of text without relying on any form of training data. These models date back to the early work on latent semantic analysis (LSA) by Deerwester, Dumais, Landauer, Furnas, and Harshman (1990), who proposed applying SVD to decompose a document-term matrix to uncover the associations between terms and concepts in the data. In basic terms, a topic model consists of *k* topics, each represented by a ranked list of strongly-associated terms (often referred to as a "topic descriptor"). Each document in the corpus can also be associated with one or more of these topics to varying degrees.

Considerable research on topic modeling has focused on the use of probabilistic methods, where a topic is viewed as a probability distribution over words, with documents being mixtures of topics (Steyvers & Griffiths, 2007). The most widely-applied probabilistic topic modeling approach has been LDA (Blei et al., 2003). Different approximation methods have been proposed for LDA inference, including variational inference and Markov chain Monte Carlo (MCMC). Such approximation algorithms can converge to different local maxima on the same data (Zhao et al., 2015). The most commonly-used implementation, provided by the *Mallet* software package (McCallum, 2002), relies on fast Gibbs sampling, where the initial state is determined by a user-specified random seed.

Alternative algorithms, such as Non-negative Matrix Factorization (Lee & Seung, 1999), have also been effective in discovering topics in text corpora (Arora et al., 2012; Kuang et al., 2015). NMF is an unsupervised approach for reducing the dimensionality of non-negative matrices. When working with a document-term matrix **A**, the goal of NMF is to approximate this matrix as the product of two non-negative factors **W** and **H**, each with *k* dimensions. The rows of the factor **H** can be interpreted as *k* topics, defined by non-negative weights for each of the *m* terms in the corpus vocabulary. Ordering each row provides a topic descriptor, in the form of a ranking of the terms relative to the corresponding topic. The columns in the matrix **W** provide membership weights for all documents with respect to each of the *k* topics. One of the advantages of NMF over traditional LDA methods is that there are fewer parameter choices involved in the modeling process, while

it also has a tendency to identify more coherent topics than LDA (O'Callaghan, Greene, Carthy, & Cunningham, 2015).

NMF is commonly initialized by assigning random non-negative weights to the entries in the factors **W** and **H**. By applying an optimization process, such as alternating least squares (Lin, 2007), the factors are iteratively improved to reduce the approximation error until a local minimum is reached. As a result, the values in the initial pair of factors will have a significant impact on the values in the final factors (*i.e.* the topic-term and topic-document weights), even after a large number of iterations have been performed. Alternative initialization schemes for NMF have focused on increasing the accuracy of the final factors by using a more structured process, such as seeding using a prior clustering algorithm (Wild, Curry, & Dougherty, 2004). Another approach, Non-negative Double Singular Value Decomposition (NNDSVD) (Boutsidis & Gallopoulos, 2008), chooses initial factors based on a sparse SVD approximation of the original data matrix. This has been shown to be particularly effective on sparse data, such as text (O'Callaghan et al., 2015). In its basic form, NNDSVD contains no stochastic element and should technically converge to the same pair of factors each time, although this depends on the underlying SVD implementation being used.

2.2. Stability in cluster analysis

Partitional clustering algorithms, such as *k*-means and *k*-medoids, have an inherent stability problem. That is to say, if we run the same algorithm on the same data or data drawn from the same source repeatedly, we frequently achieve different results between each run. This variation can either be due to poor random seeds leading to convergence to different local minima (Pena et al., 1999), or as a result of perturbations in the data (Ben-Hur, Elisseeff, & Guyon, 2002).

One widely-adopted approach for dealing with the issue is to adopt a better cluster initialization strategy that is either fully deterministic or at least produces less variation than random initialization, while simultaneously yielding more useful clusterings. A popular initialization approach proposed by Arthur and Vassilvitskii (2007), referred to as *k-means++*, involves choosing an initial seed item at random as the first cluster center and then choosing each subsequent cluster center with a probability proportional to its squared distance from the items nearest existing cluster centers. To further improve the resulting clustering, this process can be repeated for several different initial seed items. While this strategy is not deterministic, it does tend to yield more consistent results across multiple runs. Researchers have also proposed fully deterministic strategies, where initial cluster centers are determined based on embedding methods such as PCA (Su & Dy, 2004), or coming from the prior application of another algorithm such as hierarchical clustering (Celebi & Kingravi, 2012).

2.3. Ensemble methods

An alternative strategy for reducing instability in unsupervised learning is to use ensemble clustering techniques, which are based on the premise that combining large, diverse sets of clusterings can produce a more stable and accurate solution (Strehl & Ghosh, 2002). Ensemble approaches are usually divided into two different stages. Firstly, a collection of base clusterings are generated (*i.e.* the ensemble members), typically by repeatedly applying an algorithm such as *k*-means with random initialization to the full dataset or to random samples of the data (Minaei-Bidgoli, Topchy, & Punch, 2004). Secondly, an integration function is applied to combine the base clusterings into a single consensus clustering. One of the most common integration strategies utilized is to leverage information from the ensemble regarding the

---

[1] See https://github.com/derekgreene/topic-ensemble.

level of "co-association" between all pairs of items. The underlying idea behind this is that items that are frequently assigned together in different clusterings will naturally belong to the same underlying group (Strehl & Ghosh, 2002). The resulting consensus clustering represents an approximation of the "average" clustering from among the ensemble members. Hadjitodorov, Kuncheva, and Todorova (2006) demonstrated a trade-off between diversity and quality in cluster ensembles, and proposed a number of measures to quantify diversity for such ensembles.

Work regarding the optimality and consistency of solutions produced by clustering and biclustering algorithms has been previously carried out in other domains (Bertoni & Valentini, 2005; Pio, Ceci, Malerba, & D'Elia, 2015). However, in the general area of matrix factorization, there has been only some initial work on the use of ensemble approaches. This includes using a hierarchical scheme to combine multiple factorizations in the study of protein networks (Greene, Cagney, Krogan, & Cunningham, 2008) and the generation of ensembles of factorizations via a boosting-like approach (Suh, Choo, Lee, & Reddy, 2016). Recent work has also looked at using stability as a means of identifying an appropriate number of topics in a given corpus when applying NMF to text data (Greene, O'Callaghan, & Cunningham, 2014). However, these studies have not investigated the extent of the problems introduced by instability in the context of topic modeling.

3. Stability of topic modeling

In this section we introduce three new measures for assessing the stability of a collection of topic models, and use these to demonstrate how standard NMF and LDA approaches can be prone to produce unstable results when applied to text corpora.

3.1. Overview

As discussed in Section 2, standard implementations of topic modeling approaches, such as LDA and NMF, commonly employ stochastic initialization prior to optimization. As a result, the models they produce can vary quite considerably between different runs. Regardless of whether we are applying probabilistic or non-probabilistic algorithms, we can observe that this variation manifests itself in two ways: in relation to term-topic associations, or document-topic associations. In the former, the ranking of the top terms that describe a topic can change significantly between runs. In the latter, documents may be strongly associated with a given topic in one run, but may be more closely associated with an alternative topic in another run. In more extreme cases, a consequence of both manifestations is that topics can "appear" or "disappear" across different runs of the algorithm. This presents a challenge for domain experts who seek to gain a reliable insight into a particular corpus of documents. Depending on the topics resulting from a given algorithm run, their interpretation of the data may change considerably. However, the implications of this variation are rarely discussed in the topic modeling literature, particularly in the context of matrix factorization.

3.2. Measuring stability

To actually quantify the level of stability/instability present in a collection of topic models $\{\mathcal{M}_1, \ldots, \mathcal{M}_r\}$ generated over r runs on the same corpus, we propose three measures which reflect both aspects of topic model stability as described above. These measures are general in the sense that they can be applied to models generated using either probabilistic or matrix factorization algorithms.

3.2.1. Descriptor set difference

If the topics present in two topic models are similar, we should naturally expect that the prominent terms appearing in the topic descriptors in both models will be similar. Formally, if we represent each topic in a single model $\mathcal{M}_i$ with a number of top-ranked terms t, we can calculate the *descriptor set* as the union of top terms across all k topics, which we denote T_i. By measuring the symmetric difference between the descriptor sets for two different models, we can broadly gauge the similarity of the two models. This is useful as we can capture the variance at the descriptor level as terms may appear and disappear between runs. Formally, given two topic models $\mathcal{M}_i$ and $\mathcal{M}_j$, each containing k topics represented by their top t terms, we calculate the *descriptor set difference* as:

$$DSD(\mathcal{M}_i, \mathcal{M}_j) = \frac{T_i \triangle T_j}{t \times k} \tag{1}$$

A value of 0 indicates identical descriptor sets (*i.e.* no difference), while a value of 1 indicates that the topic descriptors for the two models share no common terms at all. Given a collection of r topic models, we can calculate the Average Descriptor Set Difference (ADSD):

$$ADSD = \frac{1}{r \times (r-1)} \sum_{i,j,i \neq j}^{r} DSD(\mathcal{M}_i, \mathcal{M}_j) \tag{2}$$

This produces a value $\in [0, 1]$, where a value closer to 0 for Eq. (2) is indicative of a more stable collection of models.

3.2.2. Topic-term stability

While the ADSD gives an overall measure of the difference between two models, it does not account for cases where topics are "mixed" across different runs of the algorithm (*i.e.* the same terms appear in different topics across different runs). Therefore, we propose a measure that compares the similarity between two topic models based on a pairwise matching process at the topic level. This is important as topics may appear and disappear between different runs and also helps to capture the variance at the individual topic level.

First, given a pair of individual topics represented by their top t terms, we can measure the similarity between them based on the Jaccard Index:

$$Jac(R_i, R_j) = \frac{\left| R_i \cap R_j \right|}{\left| R_i \cup R_j \right|} \tag{3}$$

where R_i denotes the top t ranked terms for the ith topic (*i.e.* its topic descriptor). We can use the above to build a measure of the agreement between two complete topic models, each containing k topics. We construct a $k \times k$ similarity matrix $\mathbf{S}$, such that the entry S_{xy} indicates the agreement between the xth topic in the first model and the yth topic in the second model, as calculated using Eq. (3). We then find the best match between the rows and columns of $\mathbf{S}$ (*i.e.* the topics from the first model and the second model). The optimal permutation π may be found in $O(k^3)$ time by solving the minimal weight bipartite matching problem using the Hungarian method (Kuhn, 1955). From this, we can produce a Term Stability (TS) score:

$$TS(\mathcal{M}_i, \mathcal{M}_j) = \frac{1}{k} \sum_{x=1}^{k} Jac(R_{ix}, \pi(R_{ix})) \tag{4}$$

where $\pi(R_{ix})$ denotes the topic in model $\mathcal{M}_j$ matched to R_{ix} in model $\mathcal{M}_i$ by the permutation π. Values for the above take the range [0, 1], where a comparison between two identical k-way topic models will result in a score of 1.

For a collection of r topic models $\{\mathcal{M}_1, \ldots, \mathcal{M}_r\}$, we can calculate the Average Term Stability (ATS):

$$ATS = \frac{1}{r \times (r-1)} \sum_{i,j,i \neq j}^{r} TS(\mathcal{M}_i, \mathcal{M}_j) \tag{5}$$

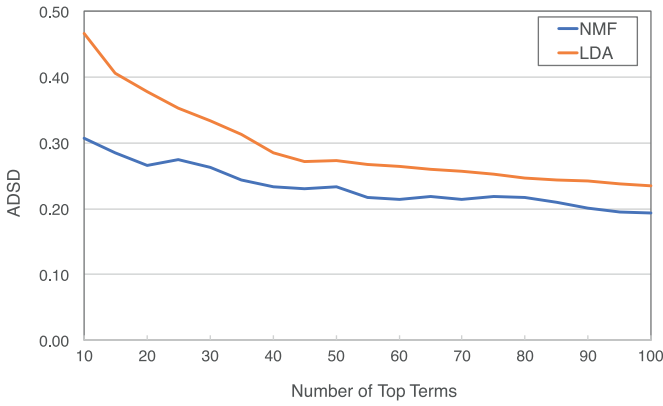where a score of 1 indicates that all pairs of topic descriptors

Fig. 1. Plot of the Average Descriptor Set Difference (ADSD) for 100 runs of NMF and LDA on the *nytimes-2003* corpus, for increasing numbers of top terms representing each topic.

matched together across the *r* runs contain the same top *t* terms.

3.2.3. Partition stability

The second manifestation of topic model instability relates to document-topic associations. To measure the extent to which the associations between a document and one or more topics varies across different runs, for each run we can look at the dominant topic for every document. That is, we convert the document-topic associations (the probabilities in the case of LDA or the **W** factor weights in the case of NMF) into a disjoint partition by taking the maximum value for each document. We can then compare the similarity between the partitions generated in two runs using standard clustering agreement measures. Utilizing this information allows us to observe if the dominant topic for each document changes frequently between runs. One widely-used such measure is Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002), which quantifies the level of agreement between two partitions P_i and P_j:

$$NMI(P_i, P_j) = \frac{I(P_i, P_j)}{\sqrt{H(P_i)H(P_j)}} \quad (6)$$

where $I(P_i, P_j)$ is the mutual information between the assignments in the two partitions and $H(P_i)$ is the entropy of the assignments in P_i alone.

We can compute the overall level of agreement between a set of *r* partitions generated by *r* runs of an algorithm on the same corpus as the mean Pairwise Normalized Mutual Information (PNMI) for all pairs:

$$PNMI = \frac{1}{r \times (r-1)} \sum_{i,j,i \neq j}^{r} NMI(P_i, P_j) \quad (7)$$

where P_i is the partition produced from the document-topic associations in model $\mathcal{M}_i$. If the partitions across all models are identical, PNMI will yield a value of 1.

3.3. Examples of instability

We now provide examples of how the measures proposed above can reflect the problem of instability, using a corpus of news articles from the New York Times as an example (the *nytimes-2003* corpus described later in Section 5).

Firstly, to illustrate the issue of term instability, we consider topic models generated for *r* = 100 runs of randomly-initialized NMF and LDA, with a fixed number of topics *k* = 7 (corresponding to the number of annotated categories in the data). Fig. 1 shows the ADSD scores for each algorithm, as the number of top terms

Table 1
Top-ranked terms for topics related to "sport", generated across five runs of randomly-initialized NMF and LDA applied to a news corpus.

#	LDA Top 10 Terms
1	game, team, season, play, coach, games, points, players, against, football
2	game, season, team, coach, play, games points, league, football, players
3	game, season, coach, team, football, league, giants, play, jets, players
4	game, season, team, yankees, games, play, mets, nets, left, league
5	game, team, season, play, games, players, coach, yankees, time, against

#	NMF Top 10 Terms
1	game, season, team, yankees, games, nets, play, points, players, coach
2	game, team, season, nets, points, games, coach, play, knicks, players
3	game, team, season, nets, points, games, play, coach, knicks, giants
4	game, nets, team, season, coach, points, knicks, jets, giants, play
5	game, team, season, nets, points, games, play, knicks, coach, kidd
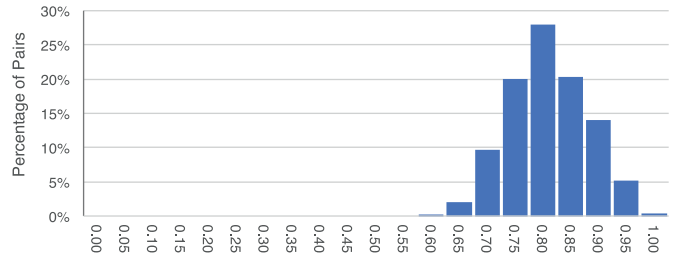
(a) NMF
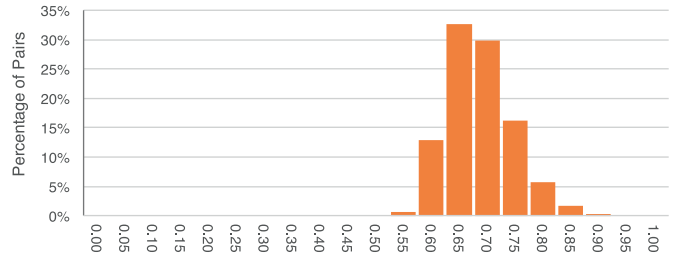
(b) LDA

Fig. 2. Distribution of pairwise NMI agreement scores for document partitions resulting from 100 runs of (a) NMF, (b) LDA on the *nytimes-2003* corpus.

in the topic descriptors increases from 10 to 100. We can observe that, even with this relatively relaxed measure, there exists substantial variation in the terms appearing in the models for both algorithms. If we represent each topic using 10 terms, the ADSD score is as high as 0.47 for LDA and 0.31 for NMF. Even when we extend the topic descriptors to contain 100 terms, which we might expect to capture the bulk of the key terms for the topics in this corpus, the mean difference across runs is 0.23 and 0.19 respectively.

We can explore this term instability further by inspecting the topic-term stability of the topics. As an example, Table 1 refers to five separate runs of a related topic (corresponding to the category "sport" in the ground truth for this corpus) for NMF and LDA. For both algorithms it is clear that the ordering of the top terms for this topic can change considerably, and it is also possible for terms to completely disappear between different runs.

Using the same corpus, we can also explore the partition stability afforded by both NMF and LDA. Fig. 2 plots the distributions for the NMI agreement scores between all pairs of partitions corresponding to the set of 100 models produced by each algorithm. Two partitions with identical document assignments would yield an NMI score of 1. However, only 0.4% of all pairs of partitions achieve this for NMF. In the case of LDA, of the 4950 unique pairs of partitions, none achieve perfect agreement and only 0.3%
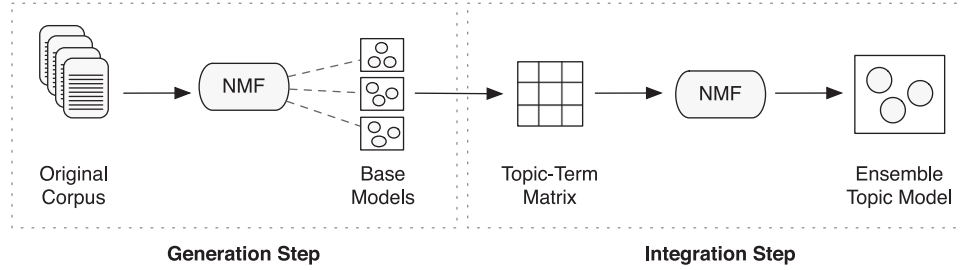
Fig. 3. Overview of a general ensemble strategy for topic modeling with matrix factorization, consisting of two steps: generation and integration.

Table 2
Top-ranked documents for topics related to "sport", generated across five runs of randomly-initialized NMF and LDA applied to a news corpus.

#	LDA Top 10 Documents
1	s4310, s5376, s4247, s5262, s6055, s1493, s3167, s5670, s4972, s6636
2	s4441, s6267, s4247, s3521, s8146, s5262, s4708, s4681, s4460, s8937
3	s0113, s3521, s4708, s1698, s5299, s4972, s8577, s8351, s5855, s6834
4	s6267, s0113, s5376, s9894, s3521, s5262, s9116, s4681, s4708, s8937
5	s6267, s0113, s9894, s4247, s5262, s4681, s2056, s4972, s8577, s6636

#	NMF Top 10 Documents
1	s5995, s6558, s3547, s9993, s5281, s8029, s2484, s5114, s1227, s2934
2	s6558, s5995, s8029, s5457, s5843, s2484, s9993, s1227, s5193, s2068
3	s8029, s6558, s5995, s5457, s5843, s2484, s1227, s9993, s5193, s2068
4	s5457, s8029, s5843, s6558, s9993, s5193, s2068, s7687, s2484, s9924
5	s8029, s6558, s5995, s5457, s5843, s2484, s1227, s9993, s5193, s2265

achieve an NMI score ≥ 0.9. When we average the agreement scores over all runs, the overall PNMI scores for the two algorithms are 0.78 and 0.66 respectively, indicating there is considerable variation in the outputs of both algorithms across these runs.

Again, manually inspecting the top-ranked documents in the topics for these models reveals the extent of the variation. Table 2 lists the identifiers of the top ten documents assigned to each of the topics related to "sport" selected from five runs of NMF and LDA in Table 1. We observe that, similar to the case of the top terms, the ordering of documents is also subject to the same inherent instability.

3.4. Redundant stability

While the examples above demonstrate that the production of robust, reliable topic models is important, it is also necessary to emphasize that stability should not be the sole requirement for a useful topic modeling algorithm. As observed by Ben-Hur et al. (2002) in the context of partitional clustering, in some situations stability can simply be indicative of an algorithm's tendency to converge to a given local solution, regardless of the quality of that solution. In the context of NMF, we could initialize the factors **W** and **H** in a deterministic way with arbitrary non-negative values. However, this "redundant stability" is unlikely to provide a useful model. Therefore, in the next section we propose techniques that yield solutions that are not only stable but also accurate – *i.e.* the topics are semantically coherent and provide a useful insight into the content of the corpus.

4. Methods

We now propose ensemble methods for topic modeling via matrix factorization, which can be utilized to address the issue of stability, while also potentially producing more accurate topic models for a corpus of unstructured text.

4.1. Basic ensemble method

We apply ensemble learning for topic modeling in the form of two layers of matrix factorization. Fig. 3 shows an overview of the method, which can naturally be divided into two steps, similar to existing strategies in ensemble clustering (Strehl & Ghosh, 2002):

1. *Generation*: Create a set of *base topic models* by executing r runs of NMF applied to the same corpus, represented as a document-term matrix **A**.
2. *Integration*: Transform the base topic models to a suitable intermediate representation, and apply a final run of NMF to produce a single *ensemble topic model*, which represents the final output of the method.

We now discuss both of these steps in more detail.

4.1.1. Ensemble generation

Unsupervised ensemble procedures typically seek to encourage diversity with a view to improving the quality of the information available in the integration phase (Topchy, Jain, & Punch, 2005). Therefore, we create a diverse set of r base topic models (*i.e.* the topic term descriptors and document assignments will differ from one base model to another). Here we encourage diversity by relying on the inherent instability of NMF with random initialization – we generate each base model by populating the factors **W** and **H** with values based on a different random seed, and then applying NMF to **A**. In each case we use a fixed pre-specified value for the number of topics k. After each run, the **H** factor from the base topic model (*i.e.* the topic-term weight matrix) is stored for later use.

4.1.2. Ensemble integration

Once we have generated a collection of r factorizations, in the second step we create a new representation of our corpus in the form of a topic-term matrix **M**. The matrix is created by stacking the transpose of each **H** factor generated in the first step, as illustrated in Fig. 4. Here each factor consists of k topics $\{t_1, \ldots, t_k\}$ and m terms $\{w_1, \ldots, w_m\}$. We construct this topic-term matrix as we may often expect to see similar topics appearing between different runs. However, they may not be identical with respect to their terms, and we wish to leverage this variance. It is important to note that this process of combining the factors is order independent. This results in a matrix **M** where each row corresponds to a topic from one of the base topic models, and each column is a term from the original corpus. Each entry M_{ij} holds the weight of association for term i in relation to a single topic from a base model.

Once we have created **M**, we apply the second layer of NMF to this matrix to produce the final ensemble topic model. The reasoning behind applying NMF a second time to these topic descriptors is that they explicitly capture the variance between the base topic models. To improve the quality of the resulting topics, we generate initial factors using NNDSVD initialization (Boutsidis & Gallopoulos, 2008). As an input parameter to NMF, we specify a final
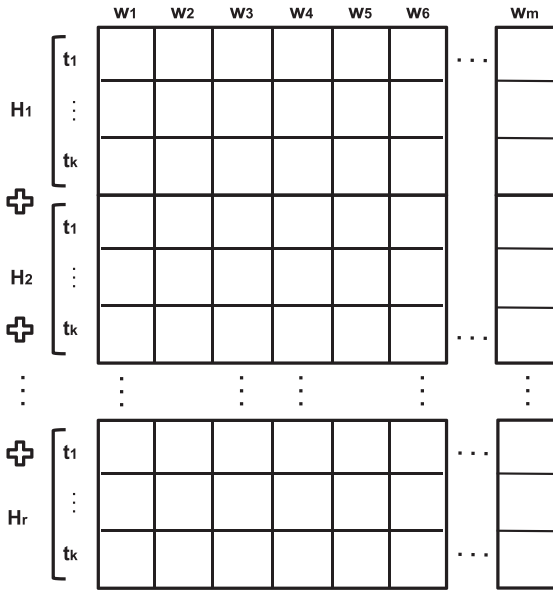
Fig. 4. Illustration of the stacking process where the **H** factors from *r* topic models, each consisting of *k* topics and *m* terms, are combined to create a single topic-term matrix.

number of k' topics, which is typically set to be the same as the value k used in the generation step. The resulting **H** factor provides weights for the terms for each of the k' ensemble topics – the top-ranked terms in each column can be used as descriptors for a topic. To produce weights for the original documents in our corpus, we can "fold" the documents into the ensemble model by applying a projection to the document-term matrix **A**:

$$D = A \cdot H^T$$

Each row of **D** now corresponds to a document, with columns corresponding to the k' ensemble topics. An entry D_{ij} indicates the strength of association of document i in ensemble topic j.

4.2. K-Fold ensemble method

While the basic ensemble generation approach described in Section 4.1.1 does yield a diverse set of base topic models, the use of random initialization means that some of these models will correspond to poor local minima with low accuracy. Furthermore, given the number of possible initial factors that could be generated in this way, there is still potential for several runs of the complete ensemble process to yield somewhat different final results. Therefore, we consider the use of improved initialization to generate more accurate base models, while also using a more structured strategy to create the models in order to reduce variability. This strategy is based on traditional k-fold cross-validation as performed in evaluation in supervised learning.

In our case, we randomly divide the corpus of documents into f folds of equal size. Each of the f folds is excluded in turn, and we apply NMF with NNDSVD initialization to the documents from the remaining $(f-1)$ folds, yielding f models. To reduce variability, we repeat the process for p rounds using different splits of the data, yielding a total of $p \times f$ topic models, each generated on a large subsample of the corpus. This collection of base topic models is then integrated as described in Section 4.1.2 to produce a final topic model. A full summary of this approach is given in Fig. 5.

1. Construct full document-term matrix **A** for the corpus.
2. For p rounds:
   1. Randomly divide corpus into f folds.
   2. For each of the f folds:
      (a) Exclude the current fold.
      (b) Apply NMF with NNDSVD initialization to documents in **A** from the other $(f-1)$ folds to generate k topics.
3. From the resulting $p \times f$ models, construct a topic-term matrix **M** by stacking the transpose of each **H** factor.
4. Apply NMF with NNDSVD initialization to **M** to generate k' topics, where the factor **H** gives the final topic-term associations.
5. Compute $D = A \cdot H^T$ to find the final document-term associations.

Fig. 5. Summary of K-Fold ensemble topic modeling method.

5. Evaluation

In this section we comprehensively assess the problem of instability in topic modeling for standard NMF and LDA approaches on a diverse collection of corpora, and examine the extent to which superior initialization and ensemble methods can improve the stability of NMF-based approaches, while also yielding accurate models.

5.1. Datasets

For our experiments, we use a diverse set of ten corpora, including both high-quality long texts and user-generated content. All of these corpora have human annotated "ground truth" topical categories, allowing us to evaluate model accuracy. Six of these datasets consist of news articles from individual mainstream news sources (BBC, The New York Times, The Guardian, and The Irish Times), categorized by subject matter. Two more datasets consist of pages from a 2014 Wikipedia dump, categorized by their associated WikiProject. These eight datasets were previously used for topic modeling evaluations (Greene, O'Callaghan, & Cunningham, 2014). We also include the popular *20-newsgroups* dataset, where the ground truth categories correspond to individual newsgroups (*e.g.* "comp.graphics", "comp.windows.x", "rec.autos").

To evaluate performance on social media data, we include a newly-collected corpus in our experiments, known as the *20-topics* dataset, which consists of 4,170,382 tweets from 1200 prominent Twitter accounts. These accounts have been manually assigned to 20 different categories (*e.g.* "aviation", "health", "tech"). Each document in the corpus corresponds to the concatenation of the tweets posted by a single user for a given week during the period March 2015 to February 2016. The corpus contains 40,498 such "user documents". A detailed summary of all corpora used in our experiments is provided in Table 3.

5.2. Experimental setup

When pre-processing the corpora, terms appearing in < 20 documents are filtered. We use a single list of common English stopwords for all datasets. LDA operates on bag-of-words text representations, and so was applied to the raw frequency values. For NMF, the same documents were transformed to log-based Term Frequency-Inverse Document Frequency (TF-IDF) vectors, and document length normalization was subsequently applied to produce the final document-term matrix.

Table 3
Details of the ten corpora used in our experiments, including the total number of documents n, number of terms m, and number of categories $\hat{k}$ in the associated "ground truth" annotations.

Corpus	n	m	$\hat{k}$	Description
bbc	2225	3121	5	General news articles from the BBC from 2003.
bbc-sport	737	969	5	Sports news articles from the BBC from 2003.
guardian-2013	6520	10,801	6	Corpus of news articles published by The Guardian during 2013.
irishtimes-2013	3246	4832	7	Corpus of news articles published by The Irish Times during 2013.
nytimes-1999	9551	12,987	4	A subset of the New York Times Annotated Corpus from 1999.
nytimes-2003	11,527	15,001	7	A subset of the New York Times Annotated Corpus from 2003.
wikipedia-high	5738	17,311	6	Subset of 2014 Wikipedia dump, where articles are assigned labels based on their high level WikiProject.
wikipedia-low	4986	15,441	10	Subset of 2014 Wikipedia dump, where articles are labeled with fine-grained WikiProject sub-groups.
20-newsgroups	18,662	9954	20	Collection of posts from 20 different internet newsgroups.
20-topics	40,498	32,464	20	Tweets from user accounts associated with 20 different topical areas.

In our experiments, we compare five different topic modeling approaches:

1. Standard LDA with random seeding, using the popular *Mallet* implementation with Gibbs sampling (McCallum, 2002).
2. NMF with random initialization, using the fast alternating least squares variant proposed by Lin (2007) and provided by the *sckit-learn* toolkit (Pedregosa et al., 2011).
3. NMF with non-random NNDSVD initialization, also implemented in *sckit-learn*.
4. Basic ensemble topic modeling for matrix factorization with random initialization, as described in Section 4.1.
5. K-Fold ensemble topic modeling for matrix factorization combined with improved initialization, as described in Section 4.2.

For these approaches, there are a number of common and distinct parameters which need to be specified:

Common parameters: For all approaches, the number of topics k is set to correspond to the number of ground truth categories for each dataset.

NMF parameters: For NMF with both random and NNDSVD initialization, the maximum number of iterations is set to 100 by default. For the random case, a different random seed is used for each run to populate values in the initial factors **W** and **H**. This process is repeated for $r = 100$ runs.

LDA parameters: The LDA algorithm has two additional hyperparameters. We use the Mallet default values, with $\alpha = 5$ and $\beta = 0.01$. The maximum number of iterations is set to 1000. For each run, a different random seed is used to initialize the Gibbs sampling process. This process is repeated for $r = 100$ runs.

Basic ensemble: For the first ensemble approach, we integrate a collection of 100 members, generated via random initialization. The final number of topics k' is set to be the same as the number of ground truth categories for each dataset. This entire process is repeated 20 times to allow us to assess stability.

K-Fold ensemble: For the second ensemble approach, we apply $p = 10$ rounds of $f = 10$ folds, thus also yielding a collection of 100 ensemble members for integration, with k' determined as above. Again this entire process is repeated 20 times.

5.3. Model stability

To assess the stability of a collection of models generated by each algorithm, we use the term-based measures **ADSD** (Eq. (2)) and **ATS** (Eq. (5)) using the top $t = 10$ terms for each topic, and the document-level **PNMI** measure (Eq. (7)). Results for these measures are shown in Tables 4, 5, and 6 respectively. Across all three measures, we observe that the NNDSVD NMF and K-Fold approaches

clearly yield the most stable results. Both of these methods produce models with perfect stability for the majority of our datasets - *i.e.* they yield models in which the topic-term and document-topic associations remain the same. As expected, the randomly-initialized approaches perform the worst due to their inherent instability caused by stochastic elements as can be identified by their standard deviation scores. While the basic ensemble approach yields high stability for the smaller corpora, we do see some variation between different runs at the term and document level for the larger corpora. Here, the random initialization used when generating the ensemble members is still leading to variation in the results at the final ensemble integration phase, even with 100 ensemble members. However, the more structured nature of the generation phase for the K-Fold approach effectively negates this problem.

It is interesting to observe that, for the *20-newsgroups* and *20-topics* datasets, which contain noisier user-generated content and a larger number of underlying topics, the K-Fold ensemble approach yields higher levels of stability than NNDSVD-initialized NMF. This suggests that combining the subsampling element of the ensemble process with a structured NNDSVD initialization produces a more reliable solution.

It is important to note that the widely-used implementation of NNDSVD provided by the *sckit-learn* toolkit, as used in our experiments, relies on an approximate truncated singular value decomposition method involving randomization, in order to make it applicable to large data matrices. While the resulting decompositions are often identical, this is not always the case. Computing a full SVD would eliminate the instability, with the trade-off that the running time requirements for decomposing a large, high-dimensional document-term matrix would increase dramatically.

5.4. Model quality

The primary focus of our work is on model stability. But as noted in Section 3.4, stability without meaningful and coherent topics is unlikely to be useful. Therefore we consider the quality of the models produced by each of the five methods, in terms of accuracy and coherence.

Partition accuracy: In either the case of NMF or LDA, we can convert a model's document-topic associations into a disjoint partition. We can then compare this partition with the corresponding disjoint "ground truth" categories for each corpus using NMI (Strehl & Ghosh, 2002).

Topic coherence: Coherence refers to the overall quality and the semantic relatedness of the terms appearing in a topic descriptor. While a range of measures have been proposed in the literature, we employ a widely-used measure, Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009), which uses term co-occurrence counts from the full corpus to measure the average coherence of the topics in a given

Table 4

Model stability — comparison of Average Descriptor Set Difference (ADSD) scores for five topic modeling approaches.

Corpus	LDA	NMF	NNDSVD	Ensemble	K-Fold
bbc	0.14 ± 0.15	0.15 ± 0.25	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
bbc-sport	**0.29 ± 0.14**	0.21 ± 0.21	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
guardian-2013	0.15 ± 0.16	0.18 ± 0.20	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
irishtimes-2013	0.35 ± 0.13	0.42 ± 0.22	**0.00 ± 0.00**	0.01 ± 0.01	0.01 ± 0.01
nytimes-1999	0.23 ± 0.18	0.36 ± 0.22	**0.00 ± 0.00**	0.21 ± 0.23	**0.00 ± 0.00**
nytimes-2003	0.47 ± 0.13	0.31 ± 0.18	**0.00 ± 0.00**	0.01 ± 0.01	**0.00 ± 0.00**
wikipedia-high	0.26 ± 0.12	0.21 ± 0.13	**0.00 ± 0.00**	0.01 ± 0.01	**0.00 ± 0.00**
wikipedia-low	0.21 ± 0.06	0.12 ± 0.09	**0.00 ± 0.00**	0.06 ± 0.07	0.00 ± 0.01
20-newsgroups	0.31 ± 0.06	0.32 ± 0.09	0.18 ± 0.08	0.12 ± 0.05	**0.05 ± 0.04**
20-topics	0.23 ± 0.06	0.23 ± 0.09	0.07 ± 0.07	0.11 ± 0.06	**0.01 ± 0.01**

Table 5

Model stability — comparison of term stability scores for five topic modeling approaches, calculated using Average Term Stability (ATS).

Corpus	LDA	NMF	NNDSVD	Ensemble	K-Fold
bbc	0.86 ± 0.14	0.88 ± 0.21	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
bbc-sport	0.68 ± 0.15	0.80 ± 0.20	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
guardian-2013	0.83 ± 0.18	0.83 ± 0.19	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
irishtimes-2013	0.61 ± 0.14	0.64 ± 0.18	**1.00 ± 0.00**	0.99 ± 0.01	0.99 ± 0.01
nytimes-1999	0.65 ± 0.19	0.72 ± 0.17	**1.00 ± 0.00**	0.82 ± 0.19	**1.00 ± 0.00**
nytimes-2003	0.53 ± 0.12	0.76 ± 0.14	**1.00 ± 0.00**	0.99 ± 0.01	**1.00 ± 0.00**
wikipedia-high	0.79 ± 0.15	0.77 ± 0.13	**1.00 ± 0.00**	0.99 ± 0.01	**1.00 ± 0.00**
wikipedia-low	0.79 ± 0.10	0.88 ± 0.08	**1.00 ± 0.00**	0.94 ± 0.07	0.99 ± 0.01
20-newsgroups	0.65 ± 0.07	0.66 ± 0.09	0.81 ± 0.08	0.86 ± 0.05	**0.93 ± 0.05**
20-topics	0.69 ± 0.08	0.78 ± 0.08	0.94 ± 0.07	0.89 ± 0.06	**0.99 ± 0.01**

Table 6

Model stability — comparison of document stability scores for five topic modeling approaches, based on Pairwise Normalized Mutual Information (PNMI).

Corpus	LDA	NMF	NNDSVD	Ensemble	K-Fold
bbc	0.86 ± 0.09	0.89 ± 0.10	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
bbc-sport	0.74 ± 0.08	0.87 ± 0.09	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
guardian-2013	0.84 ± 0.07	0.88 ± 0.07	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
irishtimes-2013	0.73 ± 0.07	0.81 ± 0.07	**1.00 ± 0.00**	0.99 ± 0.00	**1.00 ± 0.00**
nytimes-1999	0.69 ± 0.09	0.67 ± 0.14	**1.00 ± 0.00**	0.83 ± 0.13	0.98 ± 0.01
nytimes-2003	0.66 ± 0.06	0.78 ± 0.07	**1.00 ± 0.00**	0.97 ± 0.01	0.99 ± 0.00
wikipedia-high	0.86 ± 0.07	0.86 ± 0.04	**1.00 ± 0.00**	0.99 ± 0.00	**1.00 ± 0.00**
wikipedia-low	0.89 ± 0.04	0.89 ± 0.03	**1.00 ± 0.00**	0.96 ± 0.04	**1.00 ± 0.00**
20-newsgroups	0.63 ± 0.02	0.69 ± 0.04	0.80 ± 0.05	0.89 ± 0.04	**0.96 ± 0.03**
20-topics	0.84 ± 0.03	0.84 ± 0.03	0.97 ± 0.03	0.97 ± 0.02	**1.00 ± 0.00**

Table 7

Model quality — comparison of topic coherence scores for five topic modeling approaches, based on Normalized Pointwise Mutual Information (NPMI).

Corpus	LDA	NMF	NNDSVD	Ensemble	K-Fold
bbc	0.09 ± 0.01	0.15 ± 0.01	**0.16 ± 0.00**	**0.16 ± 0.00**	**0.16 ± 0.00**
bbc-sport	0.11 ± 0.01	0.14 ± 0.01	0.14 ± 0.00	**0.15 ± 0.00**	0.14 ± 0.00
guardian-2013	0.11 ± 0.01	0.15 ± 0.01	**0.16 ± 0.00**	**0.16 ± 0.00**	**0.16 ± 0.00**
irishtimes-2013	0.08 ± 0.01	0.12 ± 0.01	**0.13 ± 0.00**	**0.13 ± 0.00**	**0.13 ± 0.00**
nytimes-1999	0.10 ± 0.01	**0.12 ± 0.01**	**0.12 ± 0.01**	**0.12 ± 0.01**	**0.12 ± 0.01**
nytimes-2003	0.10 ± 0.03	**0.12 ± 0.01**	**0.12 ± 0.01**	0.10 ± 0.02	**0.12 ± 0.01**
wikipedia-high	0.15 ± 0.01	0.23 ± 0.01	**0.24 ± 0.00**	0.23 ± 0.00	**0.24 ± 0.00**
wikipedia-low	0.19 ± 0.01	0.24 ± 0.01	**0.25 ± 0.00**	0.24 ± 0.00	**0.25 ± 0.00**
20-newsgroups	0.12 ± 0.01	0.15 ± 0.01	**0.16 ± 0.01**	0.15 ± 0.00	**0.16 ± 0.00**
20-topics	0.19 ± 0.01	**0.30 ± 0.01**	**0.30 ± 0.00**	0.29 ± 0.00	0.29 ± 0.00

model, based on the top t terms in their descriptors. In our evaluations we use $t = 10$ terms.

With regards to the NPMI coherence of the topics produced, Table 7 shows that our proposed basic ensemble and K-Fold approaches perform the best. However, it should be noted that in most cases the differences in the average coherence scores are small. The most noticeable gap is between the LDA approach and the other NMF-based approaches, which may reflect the tendency of LDA to produce more generic and less semantically-coherent terms (O'Callaghan et al., 2015).

To provide a clearer measure of model quality, we next consider the quality of the five methods by evaluating the partition accuracy with respect to the ground truth labels. Table 8 shows the means and standard deviations of the NMI scores for the methods on the ten corpora. Here we see that the best-performing algorithms are NNDSVD-initialized NMF and the K-Fold ensemble approach, although this varies with the dataset. The randomly-initialized

Table 8
Model quality — comparison of document partition accuracy scores for five topic modeling approaches, based on Normalized Mutual Information (NMI).

Corpus	LDA	NMF	NNDSVD	Ensemble	K-Fold
bbc	0.81 ± 0.06	0.79 ± 0.04	**0.82 ± 0.00**	0.79 ± 0.00	0.80 ± 0.00
bbc-sport	0.68 ± 0.05	0.80 ± 0.06	0.83 ± 0.00	**0.85 ± 0.00**	**0.85 ± 0.00**
guardian-2013	0.77 ± 0.04	0.82 ± 0.04	0.83 ± 0.00	**0.84 ± 0.00**	**0.84 ± 0.00**
irishtimes-2013	0.68 ± 0.04	0.72 ± 0.04	**0.77 ± 0.00**	0.76 ± 0.00	**0.77 ± 0.00**
nytimes-1999	**0.64 ± 0.04**	0.53 ± 0.05	0.49 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
nytimes-2003	**0.60 ± 0.03**	0.58 ± 0.04	0.55 ± 0.00	0.56 ± 0.01	0.58 ± 0.00
wikipedia-high	0.69 ± 0.02	0.72 ± 0.01	0.72 ± 0.00	0.73 ± 0.00	**0.74 ± 0.00**
wikipedia-low	0.84 ± 0.02	0.87 ± 0.03	0.86 ± 0.00	**0.89 ± 0.02**	0.88 ± 0.00
20-newsgroups	0.48 ± 0.01	0.47 ± 0.02	**0.49 ± 0.01**	0.48 ± 0.01	0.48 ± 0.01
20-topics	0.84 ± 0.02	0.83 ± 0.02	0.85 ± 0.01	0.87 ± 0.01	**0.88 ± 0.00**

Table 9
Posthoc p-values based on Friedman's Aligned Rank Pairwise test for each of the four measures that were statistically significant, while using the K-Fold approach as a control algorithm. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

Measure	LDA	NMF	NNDSVD	Ensemble	K-Fold
ADSD	**** 0.00001	**** 0.00002	0.645	0.273	NA
ATS	**** 0.000002	**** 0.0001	0.713	0.250	NA
PNMI	**** 0.000002	**** 0.00007	0.576	0.304	NA
NPMI	**** 0.00001	0.111	0.921	0.452	NA

Table 10
Comparison of average running time in seconds for the five topic modeling approaches.

Corpus	LDA	NMF	NNDSVD	Ensemble	K-Fold
bbc	40.74	**0.22**	0.30	22.68	30.00
bbc-sport	33.34	**0.08**	**0.08**	7.90	8.56
guardian-2013	231.33	1.98	**1.94**	198.89	189.87
irishtimes-2013	82.75	**0.85**	0.86	86.76	84.69
nytimes-1999	370.85	**2.95**	3.74	296.18	375.97
nytimes-2003	454.36	6.11	**4.83**	613.00	538.57
wikipedia-high	667.88	5.12	**3.31**	513.68	319.96
wikipedia-low	627.28	6.42	**4.07**	644.70	380.85
20-newsgroups	229.29	**14.16**	15.15	1421.44	1527.62
20-topics	1802.16	**72.07**	115.02	7226.22	6793.77

algorithms exhibit considerable variation in the quality of the models they produce, as indicated by the standard deviation scores, and are worse on average than the ensemble and SVD-based methods, with the exception of the case where LDA is applied to the two New York Times corpora.

5.5. Statistical significance

To further investigate the differences between the algorithms, we performed a series of statistical tests on the results presented in the previous section. We carried out a non-parametric Friedman's Aligned Rank test (García, Fernández, Luengo, & Herrera, 2010) for each of the five measures previously reported (ADSD, ATS, PNMI, NPMI, and NMI) to test for the presence of statistically significant differences in the results amongst the five algorithms and across the ten datasets. These tests returned p-values of 0.000004 (ADSD), 0.000002 (ATS), 0.000003 (PNMI), 0.00002 (NPMI), and 0.118 (NMI) respectively. This indicates that statistically significant differences, at the 1% confidence level, exist in the results achieved by the different algorithms for each measure, except for partition accuracy (NMI).

To determine if there was a statistically significant difference between our proposed K-Fold algorithm and the other topic modeling approaches with respect to the four remaining measures, we performed a series of Friedman's Aligned Rank Pairwise post hoc tests (García et al., 2010), with the K-Fold approach used as a control. The results from these tests are reported in Table 9. It is interesting that there is a statistically significant difference, at the 1% confidence level, between our proposed approach and the two randomly initialized topic modeling algorithms across all measures, which along with the previously reported performance of the algorithm suggests that the K-Fold approach produces more stable and higher quality topic models. There is no statistical difference between our proposed K-Fold approach, the ensemble and NNDSVD, which may indicate that these three measures are similar due to producing more deterministic solutions and this notion is further strengthened due to their similar performance regarding the measures previously reported. It is also interesting to note that there is a statistical difference between the LDA and the K-Fold approach with regards to coherence, likely due to LDA based approaches generating topics with a lower coherence (O'Callaghan et al., 2015).

5.6. Computational expense

While the timing of algorithms can vary depending on the hardware and implementation utilized, it is useful in this case to obtain an estimate of how much longer the proposed ensemble approaches take with respect to traditional topic modeling algorithms. Each topic modeling approach was run 100 times and the average times are reported in Table 10. The experiments were carried out on a machine with 12 2.4 GHz cores and 128GB of RAM. These running times are impacted due to numerous factors, including the number of documents in the corpus, the dimensionality of the corresponding document-term matrix, and the number of topics k selected for the corpus. As expected, it is clear that the two ensemble approaches take considerably longer to run than the other algorithms. This is naturally due to the underlying nature of their generation step, where 100 iterations of NMF have to be generated. While the LDA *Mallet* implementation is widely used for topic modeling, we observe that it is considerably slower in the majority of our experiments, in fact it is frequently slower than our proposed K-Fold approach which utilizes a more structured but slower initialization step.

5.7. Discussion

So far we have discussed the two main criteria for evaluating topic modeling algorithms separately. These are the evaluation of *model quality*, by examining topic coherence and partition accuracy, and the evaluation of *model stability* by examining term stability and document stability. However, it is important to note that the output of both criteria should be considered together – our evaluations highlight that some topic modeling approaches perform well with respect to one criterion, while performing poorly with respect to the other. An example of this can be seen in the results produced by NNDSVD for the *nytimes-1999* dataset. For the term and document stability, perfect stability scores are achieved (Tables 5 and 6). However, when we take into account partition accuracy for the same dataset (Table 8), the quality of this solution is not as

good as it initially appears. The NNDSVD initialization actually performs the worst with regards to accuracy in this case, with a low NMI score of 0.49. Similarly, while randomly-initialized LDA outperforms all other approaches on both New York times corpora in terms of partition accuracy (Table 8), the corresponding stability scores are consistently poor across all measures (Tables 4–6). Following the discussion of "redundant stability" in Section 3.4, these results raise an interesting problem in that, while we strive to produce the most stable results as possible, it may also be the case that these results are of poor quality from a model quality standpoint.

Among the two newly-proposed approaches, it is interesting to observe that the basic ensemble approach does not perform as well as the K-Fold approach, even though they are based on a similar ensemble process. In the case of the former, we aim to promote diversity when generating our base ensemble members by using randomly-initialized NMF, as motivated by previous work in both supervised and unsupervised ensemble learning (Brown, Wyatt, Harris, & Yao, 2005; Kuncheva & Hadjitodorov, 2004). However, for larger datasets, the stochastic nature of this approach tends to cause the final results to contain a degree of variance across different runs of the overall ensemble. In contrast, by combining structured document subsampling with NNDSVD initialization to generate each ensemble member, the K-Fold approach exhibits very little instability across the 20 runs in our experiments, as indicated in the results in Tables 4–6. These findings correspond to those of Hadjitodorov et al. (2006), who demonstrated that a moderate level of diversity leads to useful ensembles in cluster analysis.

Overall, among the techniques considered in our experiments, the K-Fold ensemble approach produces the best models when taking into account both quality and stability. While the observed NMI scores were lower for certain datasets, it still performs better than the other methods with respect to half of our corpora. This strategy also appears to handle noisy user-generated data well, in comparison to alternative techniques.

6. Conclusions

While topic modeling methods such as LDA and NMF are widely applied in a range of domains to analyze unstructured text, researchers often do not consider the effect that random initialization has on models produced by these methods. In this paper we have demonstrated that, for both methods, this can result in significant variations in the topics produced over multiple runs over the same corpus. This effect is manifested both at the term and document level, which can potentially lead to different human interpretations of the underlying thematic structure of the data.

To address the issue of instability in the context of NMF, we have investigated the extent to which improved algorithm initialization and ensemble strategies can produce more stable models, which are almost potentially more accurate and insightful. We compared the performance of these approaches with regards to five different metrics that measure stability, accuracy, and coherence of topics. Our results indicate that a new K-Fold ensemble approach afforded the most stable and accurate set of models, although initializating NMF based on a SVD approximation of the document-term matrix can also provide a clear improvement over standard NMF and LDA methods.

One concern that arises in the application of ensemble learning techniques in general relates to scalability. While the ensemble techniques described in this paper can be naturally parallelized, there is considerable scope for reducing the computation time required to generate the ensemble. A potentially promising idea to investigate in this context relates to the concept of *snapshot ensembles* from supervised learning (Huang et al., 2017), where a single algorithm run is allowed to converge to several local minima during the optimization process, each providing a contribution to the overall ensemble. Such an approach might also be used to yield more stable topic models via matrix factorization and reduce the computational expense.

References

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models – going beyond SVD. In *Proceedings of the 53rd symposium on foundations of computer science* (pp. 1–10). IEEE.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Proceedings of the 7th Pacific symposium on biocomputing* (pp. 6–17).

Bertoni, A., & Valentini, G. (2005). Random projections for assessing gene expression cluster stability. In *Proceedings of the IEEE international joint conference on neural networks (IJCNN'05): 1* (pp. 149–154).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bouma, G. (2009). Normalized pointwise mutual information in collocation extraction. In *Proceedings of the international conference of the German society for computational linguistics and language technology. In GCSL '09.*

Boutsidis, C., & Gallopoulos, E. (2008). SVD based initialization: A head start for non-negative matrix factorization. *Pattern Recognition, 41*(4), 1350–1362.

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for k-means clustering. In *Proceedings of the 15th international conference on machine learning (ICML'98): 98* (pp. 91–99).

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion, 6*(1), 5–20.

Celebi, M. E., & Kingravi, H. A. (2012). Deterministic initialization of the *k*-means algorithm using hierarchical clustering. *International Journal of Pattern Recognition and Artificial Intelligence, 26*(07), 1250018.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41*(6), 391–407.

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences, 180*(10), 2044–2064.

Greene, D., Cagney, G., Krogan, N., & Cunningham, P. (2008). Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics, 24*(15), 1722–1728. doi:10.1093/bioinformatics/btn286.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In *Proceedings of the European conference on machine learning (ECML'14)* (pp. 498–513). Springer.

Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *Information Fusion, 7*(3), 264–275.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get M for free. *CoRR, abs/1704.00109.*

Kuang, D., Choo, J., & Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional clustering algorithms* (pp. 215–243). Springer International Publishing. doi:10.1007/978-3-319-09259-1_7.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*, 83–97.

Kuncheva, L. I., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Proceedings of the IEEE international conference on systems, man and cybernetics: 2* (pp. 1214–1219). IEEE.

Kuncheva, L. I., & Vetrov, D. P. (2006). Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(11), 1798–1808.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401.* 788–91.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Computation, 19*(10), 2756–2779.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

Minaei-Bidgoli, B., Topchy, A., & Punch, W. F. (2004). Ensembles of partitions via data resampling. In *Proceedings of the international conference on information technology: Coding and computing (ITCC'04).*

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of the annual conference of the North American chapter of the association for computational linguistics. In HLT '10* (pp. 100–108).

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications (ESWA), 42*(13), 5645–5657.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830.

Pena, J. M., Lozano, J. A., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the *k*-means algorithm. *Pattern Recognition Letters, 20*(10), 1027–1040. http://dx.doi.org/10.1016/S0167-8655(99)00069-0.

Pio, G., Ceci, M., Malerba, D., & D'Elia, D. (2015). ComiRNet: A web-based system for the analysis of miRNA-gene regulatory networks. *BMC Bioinformatics, 16*(9), S7.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424–440.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3*, 583–617.

Su, T., & Dy, J. (2004). A deterministic method for initializing *k*-means clustering. In *Proceedings of the 16th IEEE international conference on tools with artificial intelligence* (pp. 784–786). IEEE.

Suh, S., Choo, J., Lee, J., & Reddy, C. K. (2016). L-EnsNMF: Boosted local topic discovery via ensemble of nonnegative matrix factorization. In *Proceedings of the IEEE 16th international conference on data mining*.

Topchy, A., Jain, A., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(12), 1866–1881.

Wild, S., Curry, J., & Dougherty, A. (2004). Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition, 37*(11), 2217–2232.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics, 16*(13), 1.