Synthetic Dataset Generation for Online Topic Modeling

Mark Belford, Brian Mac Namee, Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland
mark.belford@insight-centre.org, brian.macnamee@ucd.ie,
derek.greene@ucd.ie

Abstract. Online topic modeling allows for the discovery of the underlying latent structure in a real time stream of data. In the evaluation of such approaches it is common that a static value for the number of topics is chosen. However, we would expect the number of topics to vary over time due to changes in the underlying structure of the data, known as concept drift and concept shift. We propose a semi-synthetic dataset generator, which can introduce concept drift and concept shift into existing annotated non-temporal datasets, via user-controlled paramaterization. This allows for the creation of multiple different artificial streams of data, where the "correct" number and composition of the topics is known at each point in time. We demonstrate how these generated datasets can be used as an evaluation strategy for online topic modeling approaches.

1 Introduction

Topic modeling is an unsupervised learning task which attempts to discover the underlying thematic structure of a document corpus. Popular approaches include probabilistic algorithms such as Latent Dirichlet Allocation [2, 19], and matrix factorization algorithms such as Non-negative Matrix Factorization [21]. Topic modeling tends to operate on static datasets where documents are not timestamped. This renders the evaluation and benchmarking of these algorithms relatively straightforward, due to the availability of many datasets which have human-annotated "ground truth" reference topics.

Online topic modeling is a variant of this task that takes into account the temporal nature of a text corpus. This often involves working with a real-time stream of data, such as that found in social media analysis and in analysis procedures associated with online journalism. In other scenarios, this task involves retrospectively working with a timestamped corpus which has previously been collected and divided into distinct *time windows*. While many sources of text naturally provide temporal metadata, we are currently unaware of any readily-available source of ground truth text data for the online topic modeling task, due to the expense and difficulty of manually annotating large temporal corpora. An associated issue is that, when applying online topic modeling approaches to real-world text streams, the number of topics in the data will naturally vary and evolve over time. However, for evaluation purposes, many existing works

assume that this number remains fixed. This is not a realistic assumption due to the expected variation in topics over time due to changes in their underlying composition, known as concept drift and concept shift [13].

To accurately benchmark new online topic modeling approaches, a quantitative approach is required to determine the extent to which these approaches can correctly identify the number and composition of topics over time. However, to achieve this, a comprehensive set of datasets is required, which provide temporal information along with ground truth topic annotations. With these requirements in mind, in this paper we propose new semi-synthetic dataset generators which can introduce concept drift and concept shift into existing static text datasets in order to create artificial data streams, where the correct number of ground truth topics at each time point is known a priori. We make a Python implementation of these generators available for further research[1].

The paper is structured as follows. In Section 2 we present related work covering existing evaluation strategies for static and online topic modeling. In Section 3 we outline our proposed methodology behind two new synthetic dataset generators, before exploring the use of a number of test generated datasets in Section 4. We present our conclusions and future work in Section 5.

2 Related-work

2.1 Topic Modeling

Topic modeling attempts to discover the underlying thematic structure within a text corpus. These models date back to the early work on latent semantic indexing [5]. In the general case, a topic model consists of k topics, each represented by a ranked list of highly-relevant terms, known as a *topic descriptor*. Each document in the corpus is also associated with one or more topics.

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, thus permitting a topic model to be considered a generative model for documents [19]. The most widely-applied probabilistic topic modeling approach is Latent Dirichlet Allocation (LDA) [2]. Alternative non-probabilistic algorithms, such as Non-negative Matrix Factorization (NMF) [12], have also been effective in discovering the underlying topics in text corpora [21]. NMF is an unsupervised approach for reducing the dimensionality of non-negative matrices. When working with a document-term matrix $\mathbf{A}$, the goal of NMF is to approximate this matrix as the product of two non-negative factors $\mathbf{W}$ and $\mathbf{H}$, each with k dimensions. The former factor encodes document-topic associations, while the latter encodes term-topic associations.

2.2 Topic Model Evaluation

There are a number of different techniques used in the evaluation of traditional topic modeling algorithms. The coherence of a topic model refers to the quality

[1] `https://github.com/MarkBelford/dataset-generator`

or human interpretability of the topics. Originally a task involving human annotators [4], automatic approaches now exist to calculate coherence scores using a variety of different metrics [3, 15, 11]. In topic modeling approaches such as NMF or LDA, the most prominent topic assigned to each document by the model, also known as the document-topic assignment, can be used to calculate the overall accuracy of the model. This document-topic partition is compared to a partition generated using the ground truth labels for each document using simple clustering agreement measures [20]. Topic modeling is similar to clustering in that the number of topics to be discovered must be specified at the beginning of the process. Certain evaluation techniques investigate the challenge of finding the optimal number of topics for a given dataset in a static context [9, 22].

2.3 Online Topic Modeling

Online topic modeling is a variant of traditional topic modeling which operates on a temporal source of text, such as that found in the analysis of social networking platforms and online news media. There are a number of online approaches for both LDA and NMF, however these vary greatly between implementation. Some approaches utilise an initial batch phase to initialize the model and afterwards update the model by considering one document at a time [1]. It is also possible to create a hybrid model using this approach by iterating between an online phase and an offline phase, which considers all of the documents seen so far to try and improve the clustering results. Other approaches update the model instead by considering mini-batches to try to reduce the noise present when only considering a single document [10]. A more intuitive approach represents batches of documents as explicit time windows which allows for the observation of how the topic model evolves over time [18]. It is also possible to apply dynamic topic modeling approaches [6] to temporally ordered static datasets to produce a form of online topic modeling output. In this case a dataset is divided into distinct time periods and traditional topic modeling approaches are applied to each. The results of these models are then combined and utilized in a second topic modeling process to produce results.

2.4 Online Topic Model Evaluation

The evaluation of online topic modeling approaches tends to make use of static annotated datasets, where the number of topics is known in advance. However, these approaches frequently assume that the number of topics is fixed and does not change over time. In other cases, authors select a high value of k in order to capture the majority of possible themes. However, this creates an interpretation problem, as many noisy and irrelevant topics may also be returned by the algorithm. These evaluation choices are understandable, given that manually annotating a real-time stream of data is costly and time-consuming. In other unsupervised tasks, such as dynamic community finding, the provision of synthetically-generated datasets with predefined temporal patterns (*e.g.* the

"birth" and "death" of communities [17]) has proven useful from an evaluation perspective [8]. This has motivated the work presented in the rest of this paper.

3 Methods

The lack of annotated ground truth corpora with temporal information is problematic when evaluating online topic modeling approaches. For instance, how can we determine whether a proposed algorithm can correctly determine the number of topics in the data at a given point in time? Therefore, in this section we explore two different ways in which the distribution of topics can vary over time, and then present corresponding methodologies used to implement synthetic dataset generators based on these variations. Through user paramaterization, we can control the characteristics of the resulting datasets and the extent to which they change over time. Both generators contain stochastic elements, so that many different datasets can potentially be produced for the same parameter values.

Given the complex structure of natural language corpora, generating realistic fully-synthetic datasets is extremely challenging. As an alternative, authors have proposed generating "semi-synthetic" datasets which are derived from existing real-world corpora [7, 13]. Therefore, as the input to each of our proposed generators, we can use any existing large document corpus that has k' ground truth annotated topics, but which does not have necessarily temporal metadata. In the case of both generators, we make use of $k \leq k'$ of these annotated topics.

Both generators also operate on the principle that a single "window" of documents represents one epoch in the overall dataset – *i.e.* the smallest time unit considered by the algorithm. Depending on the context and source of data, in practice this could range from anywhere between seconds (*e.g.* in the case of tweets) to years (*e.g.* in the case of financial reports). However, for the purpose of discussion, we refer to these generally as *time windows*.

3.1 Concept Shift Generator

Concept shift refers to the change in concept due to a sudden variation in the underlying probabilities of the topics. In the context of online news, a common example might occur where the coverage of already established news stories is reduced greatly after the death of a prominent figure, while the coverage of this latter topic increases rapidly. A visual example of this can be seen in Fig. 1.

We propose a textual data generator, embedding the idea of concept shift, which operates as follows. To commence the process, k topics from the ground truth and *window-size* number of documents from these topics are randomly selected to form the initial time window. At each subsequent time window, documents are chosen from these topics. There is also a chance that, based on a user defined probability parameter, *shift-prob*, a topic is added or removed from the model. The idea is that this event will simulate a concept shift over time. This process of generating time windows continues until the number of remaining topics reaches a minimum threshold, defined by the parameter *min-topics*.

Algorithm 1 Concept Shift Generator

Parameters

- *input*: an existing dataset with ground truth topic annotations.
- k: number of starting topics.
- *window-size*: number of documents in each time window.
- *shift-prob*: the probability of a concept shift occurring.
- *min-topics*: minimum number of topics present before ending.

Algorithm

1. Randomly select k starting topics.
2. Randomly select *window-size* documents from these starting topics.
3. Generate a new time window:
   - While the number of documents in the window is less than *window-size*
     - If concept shift is activated, randomly add or remove a topic.
     - Randomly choose a topic from those already in the model.
     - Randomly choose a document from this topic.
     - Add this document to the window.
4. Repeat from Step 3 until *min-topics* remain in the model.

An overview of the complete process is given in Algorithm 1. The output of the process is a set of time window datasets, each containing documents with ground truth topic annotations.

It is important to note that, unlike a real stream of data, we do not have access to an infinite number of documents. Depending upon the size of the original input dataset, this can lead to situations where a topic that is currently present in the model can run out of documents in the middle of generating a new time window.

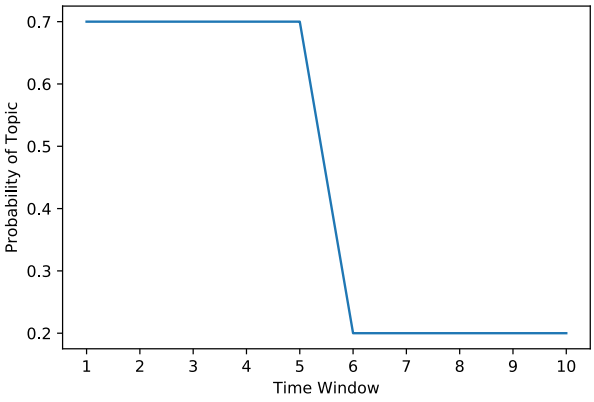


Fig. 1: Example of concept shift, where the probability of a topic appearing changes dramatically over a single time window (*i.e.* window 5 to 6).

This is handled by simply removing the topic so that it can no longer be chosen by the generator in subsequent time windows.

3.2 Concept Drift Generator

Concept drift refers to the gradual change in the underlying probabilities of topics appearing over time. An example of this is commonly seen in news media, where the coverage of an ephemeral event that is near the end of its news cycle, such as the Summer Olympics or FIFA World Cup, is gradually reduced over time. In contrast, the coverage of other newly-emergent stories may increase during this time. A simple visual example of this trend can be seen in Fig. 2.

The proposed concept drift generator (Algorithm 2) operates as follows. Firstly, k topics and *window-size* number of documents are are chosen based on randomly-assigned probabilities to form the initial window. For all remaining windows, topics are chosen based on their current probability. There is also a user-defined parameter, *drift-prob*, that determines whether a concept drift event will occur in a given window. If this occurs, then the generator will randomly choose one topic to slowly remove by decreasing its probability over a fixed number of time windows (determined by the parameter *decrease-windows*), while simultaneously choosing a new topic to slowly introduce over a fixed number of time windows (determined by *increase-windows*). This process of continues until the number of topics remaining goes below a minimum threshold (*min-topics*). The output of the process is a set of time window datasets.

However, again there is the issue that we do not have an infinite number of documents, so topics might potentially run out of documents during a drift. Unlike the previous generator we do not simply remove the topic during the middle of the drift. Instead we leave the topic in the model for the remainder of the drift and if the topic is chosen we simply ignore it. Note that this can lead

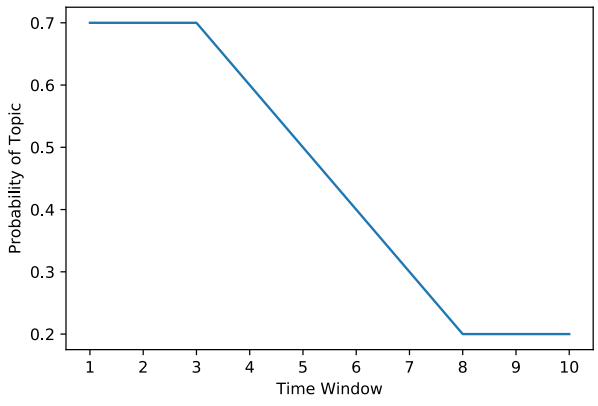


Fig. 2: Example of concept drift where the probability of a topic appearing changes gradually over a number of time windows.

Algorithm 2 Concept Drift Generator

Parameters

- *input*: an existing dataset with ground truth topic annotations.
- k: number of starting topics, must be less than the total number of topics.
- *window-size*: number of documents in each time window.
- *increase-topic*: topic to be slowly introduced by concept drift.
- *decrease-topic*: topic to be slowly removed by concept drift.
- *increase-windows*: number of windows for a topic to gradually disappear.
- *decrease-windows*: number of windows for a topic to gradually appear.
- *drift-prob*: the probability of a concept drift occurring.
- *min-topics*: minimum number of topics present before ending.

Algorithm

1. Randomly select k starting topics.
2. Randomly select *window-size* documents from these starting topics.
3. Generate a new time window:
   - While the number of documents in the window is less than *window-size*
     - If concept drift is enabled, gradually increase and decrease the probabilities of the *increase-topic* and *decrease-topic* over *increase-windows* and *decrease-windows* respectively.
     - Otherwise choose a topic from those already in the model based on their probabilities.
4. Repeat from Step 3 until *min-topics* remain in the model.

to some windows having less than *window-size* number of documents, depending upon the size of ground truth topics in the original input dataset.

4 Tests

In this section we explore sample datasets generated by our two approaches from Section 3, and demonstrate how these can be used to validate the outputs of a dynamic topic modeling approach. Note that our goal here is not to evaluate any individual topic modeling algorithm, but rather to illustrate how the proposed generators might be useful in benchmarking such algorithms.

4.1 Datasets

As our input corpus for generation, we use the popular *20-newsgroups* (20NG) collection[2] which contains approximately 20,000 Usenet postings, corresponding to roughly 1,000 posts from each of 20 different newsgroups covering a wide range of subjects (*e.g.* "comp.graphics", "comp.windows.x","rec.autos"). While this dataset has existing temporal metadata we chose not take this into consideration. We want to ensure that we artificially induce events to use as our ground

[2] Available from `http://qwone.com/~jason/20Newsgroups/`

Table 1: Summary of datasets generated from the 20NG collection, including the total number of documents n, the starting number of topics k, the range of the number of topics across all time windows, the resulting number of time windows, and the input probability parameters.

Dataset	n	k	**Range**	**Windows**	**Prob.**	**Increase/Decrease**
shift-1	4,933	5	5–6	10	0.05	NA
shift-2	9,091	10	6–10	19	0.10	NA
shift-3	12,923	15	6–15	18	0.30	NA
shift-4	18,471	20	15–20	19	0.50	NA
drift-1	3,551	5	3–5	18	0.05	10 / 5
drift-2	7,122	10	6–9	15	0.10	10 / 5
drift-3	8,766	15	6–13	17	0.30	15 / 10
drift-4	13,146	19	6–19	21	0.50	15 / 10

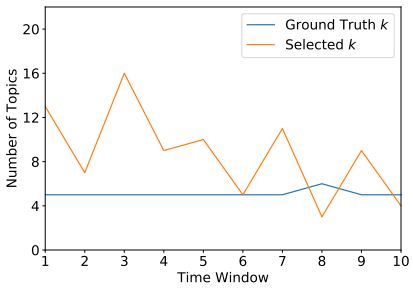
truth rather than capturing snippets of temporal events from the original data. We also choose not to utilise these timetsamps as this information is not always available and our goal is to allow the methodology to generalise to any dataset that has ground truth annotations. In the case of both generators, we make use of $k \leq k'$ of these annotated topics. We use these newsgroups as our ground truth topics. To illustrate the use of our generators, we generated four datasets which exhibit concept shift and four datasets that exhibit concept drift, using a variety of different parameter choices. A summary of the parameters and characteristics of these datasets is provided in Table 1. We observe that these sample datasets vary considerably in terms of their size, number of topics, and number of time windows. Note that the number of time windows produced by the generators is a function of the input parameters and the size of the input corpus.

4.2 Experimental Setup

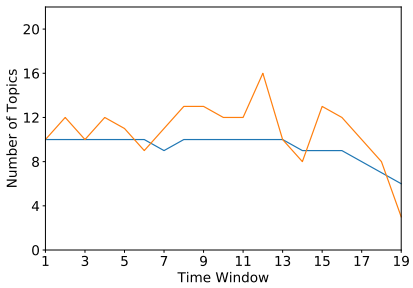
To illustrate the use of the generated datasets, we apply the window topic modeling phase from the Dynamic NMF algorithm [6], using the TC-W2V topic coherence measure [16] to select the number of topics k at each time window, as proposed by the authors. This method relies on the use of an appropriate *word2vec* word embedding model [14]. For this purpose, we construct a skip-gram *word2vec* model built on the complete 20NG corpus, with vectors of size 800 dimensions. In our experiments, we consider the range 3–20 as candidate values for k, and select the value of k with the highest coherence score.

4.3 Results and Discussion

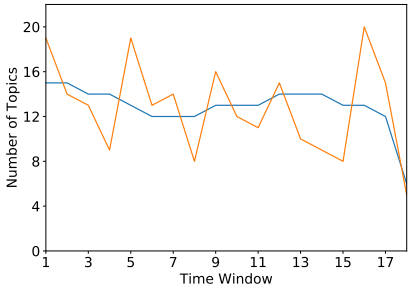
We now illustrate how our proposed generator can produce datasets that can be used for online model evaluation. Again it is important note that the performance of the approaches being applied here is not our main focus, but rather the provision of synthetic datasets that can facilitate the more robust evaluation of online topic modeling algorithms.



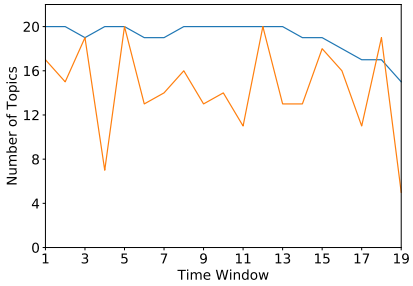
(a) *shift-1* dataset.



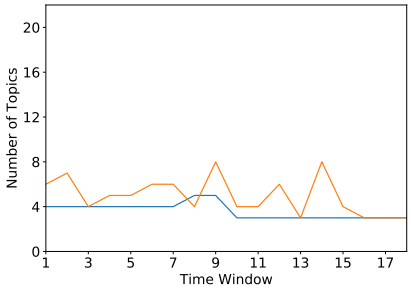
(b) *shift-2* dataset.



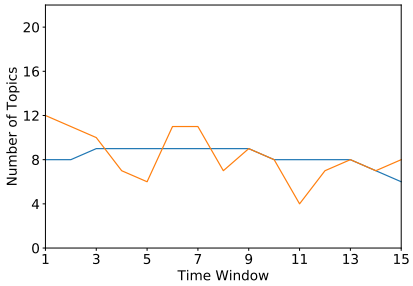
(c) *shift-3* dataset.



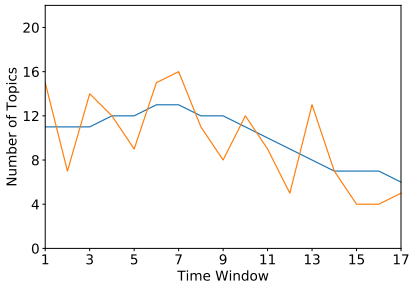
(d) *shift-4* dataset.



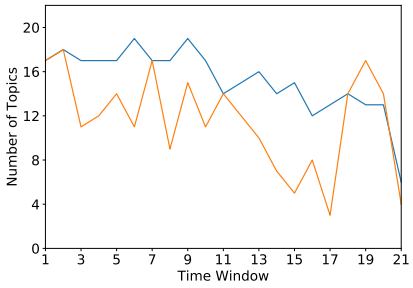
(e) *drift-1* dataset.



(f) *drift-2* dataset.



(g) *drift-3* dataset.



(h) *drift-4* dataset.

Fig. 3: Comparison of number of ground truth topics and number of topics k identified by the dynamic topic modeling approach for each time window.

Firstly, the sample generated datasets allow us to assess the extent to which the coherence-based model selection approach for NMF correctly identifies the number of topics in each time window, by comparing its selections with the number of ground truth topics in the data. Fig. 3 shows comparisons for each of the eight datasets. For many of the datasets, the selected values of k broadly follow the trend in the ground truth (where either a concept shift or drift is occurring over time), and this is most strongly seen in the concept drift dataset, *drift-4*, although we see considerable variation at individual time points. However, for the smallest concept shift dataset, *shift-1*, we see a much poorer correspondence with the ground truth when evaluating this dynamic approach. The provision of the "correct" number of topics in the ground truth potentially allows researchers to develop and benchmark methods that could provide a more useful approximation of the number of topics in these datasets.

Secondly, the generated datasets allow us to evaluate the degree to which the topics being discovered by NMF over time agree with the ground truth topics, in terms of their document assignments. To assess the topic models generated at each time window, we construct a document-topic partition from the document-topic memberships produced by NMF. This partition is compared with the annotated labels for the documents for the ground truth in the corresponding time window. To perform the comparison, we can use a simple clustering agreement score such as Normalized Mutual Information (NMI) [20]. If two partitions are identical then the NMI score will be 1, while if the two partitions share no similarities at all then the score will be 0.

Table 2 summarizes the mean and range of NMI scores across all time windows for the eight generated datasets. It is interesting to see that the performance of NMF varies considerably between the datasets, with an overall maximum value of 0.653. In some cases the level of agreement is quite poor (*e.g.* the *drift-1* dataset). This suggests considerable scope for improving topic models on these generated datasets, where NMI relative to the ground truth could provide researchers with a guideline to measure the level of improvement.

Table 2: Summary of Normalized Mutual Information (NMI) scores achieved by NMF across all time windows for each generated dataset, relative to the ground truth topics in the data.

Dataset	**Mean**	**Min**	**Max**
shift-1	0.526	0.437	0.653
shift-2	0.463	0.319	0.541
shift-3	0.512	0.390	0.613
shift-4	0.469	0.417	0.522
drift-1	0.390	0.156	0.495
drift-2	0.527	0.453	0.627
drift-3	0.458	0.339	0.520
drift-4	0.433	0.334	0.533

5 Conclusions

In this paper we have proposed two methods for generating semi-synthetic dynamic text datasets from an existing static corpus, which incorporate fundamental temporal trends – concept shift and concept drift. We have demonstrated that this generator can produce datasets with a range of different characteristics, which can be used in practice to evaluate the output of online and dynamic topic modeling methods. In particular, the generator provides a mechanism to evaluate the degree to which these methods can correctly determine the number of topics at a given point in time, relative to a set of ground truth topics. Here our focus has been on modeling the evolution of thematic structure as caused by changes in the probabilities of the underlying topics appearing. However, changes in concept can also occur due to the content of topics evolving over time [13]. In future work we plan to investigate and characterize this type of concept change in a real-time stream of text data.

Acknowledgement. This research was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. Banerjee, A., Basu, S.: Topic models over text streams: A study of batch and online unsupervised learning. In: Proceedings of the 2007 SIAM International Conference on Data Mining. pp. 431–436. SIAM (2007)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Bouma, G.: Normalized Pointwise Mutual Information in Collocation Extraction. In: Proc. International Conference of the German Society for Computational Linguistics and Language Technology. GCSL '09 (2009)
4. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.M.: Reading Tea Leaves: How Humans Interpret Topic Models. In: NIPS. pp. 288–296 (2009)
5. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
6. Greene, D., Cross, J.P.: Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Political Analysis 25(1), 77–94 (2017)
7. Greene, D., Cunningham, P.: Producing a unified graph representation from multiple social network views. In: Proceedings of the 5th annual ACM web science conference. pp. 118–121. ACM (2013)
8. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'10). IEEE (2010)
9. Greene, D., O'Callaghan, D., Cunningham, P.: How many topics? stability analysis for topic models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 498–513. Springer (2014)
10. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: Advances in neural information processing systems. pp. 856–864 (2010)

11. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: EACL. pp. 530–539 (2014)
12. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–91 (1999)
13. Lindstrom, P.: Handling Concept Drift in the Context of Expensive Labels. Ph.D. thesis, Dublin Institute of Technology (2013)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
15. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic Evaluation of Topic Coherence. In: Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. HLT '10 (2010)
16. O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications (ESWA) 42(13), 5645–5657 (2015)
17. Palla, G., Barabási, A.L., Vicsek, T.: Quantifying social group evolution. Nature 446(7136), 664–667 (2007)
18. Saha, A., Sindhwani, V.: Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization. In: Proc. 5th ACM Int. Conf. Web search and data mining. pp. 693–702 (2012)
19. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of latent semantic analysis 427(7), 424–440 (2007)
20. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (December 2002)
21. Wang, Q., Cao, Z., Xu, J., Li, H.: Group matrix factorization for scalable topic modeling. In: Proc. 35th SIGIR Conf. on Research and Development in Information Retrieval. pp. 375–384. ACM (2012)
22. Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W.: A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 16(13), 1 (2015)