Expert Systems with Applications 42 (2015) 5645–5657

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An analysis of the coherence of descriptors in topic modeling

Derek O'Callaghan *, Derek Greene, Joe Carthy, Pádraig Cunningham

School of Computer Science & Informatics, University College Dublin, Ireland

ARTICLE INFO

Article history:
Available online 9 March 2015

Keywords:
Topic modeling
Topic coherence
LDA
NMF

ABSTRACT

In recent years, topic modeling has become an established method in the analysis of text corpora, with probabilistic techniques such as latent Dirichlet allocation (LDA) commonly employed for this purpose. However, it might be argued that adequate attention is often not paid to the issue of topic coherence, the semantic interpretability of the top terms usually used to describe discovered topics. Nevertheless, a number of studies have proposed measures for analyzing such coherence, where these have been largely focused on topics found by LDA, with matrix decomposition techniques such as Non-negative Matrix Factorization (NMF) being somewhat overlooked in comparison. This motivates the current work, where we compare and analyze topics found by popular variants of both NMF and LDA in multiple corpora in terms of both their coherence and associated generality, using a combination of existing and new measures, including one based on distributional semantics. Two out of three coherence measures find NMF to regularly produce more coherent topics, with higher levels of generality and redundancy observed with the LDA topic descriptors. In all cases, we observe that the associated term weighting strategy plays a major role. The results observed with NMF suggest that this may be a more suitable topic modeling method when analyzing certain corpora, such as those associated with niche or non-mainstream domains.

1. Introduction

Topic modeling is a key tool for the discovery of latent semantic structure within a variety of document collections, where probabilistic models such as latent Dirichlet allocation (LDA) have effectively become the de facto standard method employed (Blei, Ng, & Jordan, 2003). The discovered topics are usually described using their corresponding top N highest-ranking terms, for example, the top 10 most probable terms from an LDA ϕ topic distribution over terms. In the case of probabilistic topic models, a number of metrics are used to evaluate model fit, such as perplexity or held-out likelihood (Wallach, Murray, Salakhutdinov, & Mimno, 2009). At the same time, it might be argued that less attention is paid to the issue of *topic coherence*, or the semantic interpretability of the terms used to describe a particular topic, despite the observation that evaluation methods such as perplexity are often not correlated with human judgements of topic quality (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). However, a number of measures have been proposed in recent years for the measurement of coherence, based on approaches that include co-occurrence frequencies of terms within a reference corpus (Newman, Lau, Grieser, & Baldwin, 2010; Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Lau, Newman, & Baldwin, 2014) and distributional semantics (Aletras & Stevenson, 2013). The intuition is that pairs of topic descriptor terms that co-occur frequently or are close to each other within a semantic space are likely to contribute to higher levels of coherence.

Non-probabilistic methods based on matrix decomposition are also used for topic modeling, such as Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) or Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999; Arora, Ge, & Moitra, 2012). Here, topic term descriptors can be generated in a similar fashion to those of probabilistic models, for example, using the top N highest-ranked terms from an NMF topic basis vector. In our previous work, we generated topics using both LDA and NMF with two particular corpora, where a qualitative analysis of the corresponding term descriptors found the most readily-interpretable topics to be discovered by NMF (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2013). An example of the issues we encountered can be illustrated with the following topics that were discovered by LDA and NMF for the same value of k within a corpus of online news articles (described in further detail in Section 5):

* Corresponding author.
E-mail addresses: derek.o-callaghan@ucdconnect.ie (D. O'Callaghan), derek.greene@ucd.ie (D. Greene), joe.carthy@ucd.ie (J. Carthy), padraig.cunningham@ucd.ie (P. Cunningham).

http://dx.doi.org/10.1016/j.eswa.2015.02.055
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

- LDA: *iran, syria, syrian, iraq, weapon, president, war, nuclear, military, iranian*.
- NMF: *syria, syrian, weapon, chemical, assad, damascus, rebel, military, opposition, lebanon*.

At a glance, both topics appear both relevant and coherent, with no identifiable irrelevant terms, where the topics may be interpreted as being associated with the ongoing Syria conflict. A closer inspection of the terms suggests that the LDA topic is in fact a general topic about the Middle East, while the NMF topic is far more specifically concerned with Syria (including the *lebanon* term in this context), which could also be interpreted as being more coherent depending on the end user's expectations. This issue regarding the possibility for LDA to over-generalize has been raised previously by Chemudugunta, Smyth, and Steyvers (2006). However, a study by Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012) of the coherence of topics discovered by LSA, NMF and LDA within a single corpus composed of online New York Times articles from 2003 (Sandhaus, 2008), concluded that NMF produced the more incoherent topics. As our previous findings suggest that this issue is unresolved, we perform an evaluation of LDA and NMF using a range of corpora, where our two major objectives are the measurement and comparison of (1) topic coherence, and (2) topic generality. The latter is considered at two levels; the tendency for a method to generate topics containing high-frequency descriptor terms from the underlying corpus, and also the presence of terms in multiple descriptors for a particular model, signifying the existence of overlap or dependence between the topics.

To this end, we compiled six new and existing corpora containing documents that had been (manually) annotated with classes, including online news articles from the BBC, the Guardian, and the New York Times, in addition to Wikipedia project page content. A consistent set of pre-processing steps was applied to these, and topics were discovered with LDA and NMF. Although multiple variants exist for both topic modeling methods, we restricted the experiments to those that are commonly used, with popular implementations being run accordingly (McCallum, 2002; Pedregosa et al., 2011), in addition to recommended parameter values (Steyvers & Griffiths, 2006). Two out of three coherence measures, including a new measure based on word2vec (Mikolov, Chen, Corrado, & Dean, 2013) term vector similarity, find NMF to regularly produce more coherent topics, while higher levels of generality and redundancy are observed with the LDA topic descriptors. However, we observe that the associated term weighting strategy plays a major role, as modifications to both document term pre-processing (NMF) and descriptor term post-processing (LDA) can produce markedly different results. Separately, we also find that LDA produces more accurate document-topic memberships when compared with the original class annotations.

2. Related work

2.1. Topic modeling

Topic modeling is concerned with the discovery of latent semantic structure or topics within a set of documents, which can be derived from co-occurrences of words in documents (Steyvers & Griffiths, 2006). This strategy dates back to the early work on latent semantic indexing by Deerwester et al. (1990), which proposed the decomposition of term-document matrices for this purpose using Singular Value Decomposition. Probabilistic topic models have become popular in recent years, having been introduced with the Probabilistic Latent Semantic Analysis (PLSA) method of Hofmann (2001), also known as Probabilistic Latent Semantic Indexing (PLSI). Here, a topic is a probability distribution over words, with documents being mixtures of topics, thus permitting a topic model to be considered a generative model for documents (Steyvers & Griffiths, 2006). With this process, a document is generated by first sampling a topic z from the document-topic distribution θ, followed by a word w from the corresponding topic-word distribution ϕ. The extension of this model by Blei et al. (2003), known as latent Dirichlet allocation (LDA), suggested using a Dirichlet prior on θ with an associated hyperparameter α. Griffiths and Steyvers (2004) proposed also using a Dirichlet prior on ϕ, with corresponding hyperparameter β. The plate notation for this model can be found in Fig. 1.

Griffiths and Steyvers (2004) also used collapsed Gibbs sampling to indirectly estimate these distributions, by iteratively estimating the probability of assigning each word to the topics, conditioned on the current topic assignments of all other words, using count matrices of topic-word (C^{WT}) and document-topic (C^{DT}) assignments:

$$P(z_i = j | z_{-i}, w_i, d_i, .) \propto \frac{C^{WT}_{w_i,j} + \beta}{\sum_{w=1}^{W} C^{WT}_{w,j} + W\beta} \frac{C^{DT}_{d_i,j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i,t} + T\alpha} \tag{1}$$

Following this process, the distributions for sampling a word i from topic j (ϕ^j), and topic j for document d (θ^d) are estimated as:

$$\phi^j = \frac{C^{WT}_{ij} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \qquad \theta^d = \frac{C^{DT}_{dj} + \alpha}{\sum_{t=1}^{T} C^{DT}_{dt} + T\alpha} \tag{2}$$

There have been a number of additional variants of LDA proposed in recent years. However, in this paper, we are primarily concerned with the coherence of topic modeling in general, and so the discussion here is accordingly restricted to (a) popular LDA variants, and (b) those used by the topic coherence experiments described in Section 2.2. Two popular toolkits that are often used for topic modeling with LDA are *MALLET* (McCallum, 2002), which provides a fast implementation of the Gibbs sampling method described above, and *gensim* (Řehůřek & Sojka, 2010), which implements the online variational Bayes method of Hoffman, Blei, and Bach (2010). The motivation for the latter method was the application of LDA to data streams or large datasets. In addition to the method implementations provided by MALLET and gensim, other prominent methods featuring in the topic coherence experiments that have not been discussed so far include the Correlated Topic Model (CTM) of Blei and Lafferty (2006), which attempts to directly model correlation between the latent topics themselves, and the Pólya Urn method proposed by Mimno et al. (2011), which extended Gibbs sampling to incorporate information used in the corresponding coherence metric.

Non-negative Matrix Factorization (NMF) is a technique for decomposing a non-negative matrix $V \in \mathbb{R}$ into two non-negative factors W and H, where $V \approx WH$ (Lee & Seung, 1999). Although it has been used in multiple domains, it is also applicable to topic modeling (Arora et al., 2012). In this context, V is an $n \times m$ term-document matrix, and W and H are reduced rank-k factors whose product is an approximation of V, with dimensions $W = n \times k$ and $H = k \times m$. This enables a parts-based representation, where W contains a set of k topic basis vectors, and H provides the coefficients for the additive linear combinations of these basis vectors to generate the corresponding document vectors in V. The weights in a W topic basis vector can be used to generate a topic descriptor consisting of high-ranking terms (analogous to the most probable terms in an LDA ϕ distribution), while a H vector of coefficients can be interpreted as the k topic membership weights for the corresponding document. Two common objective functions (Lee & Seung, 2001) used to generate W and H are the Euclidean squared error:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (V_{ij} - (WH)_{ij})^2 = \|V - WH\|_F^2 \tag{3}$$