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.

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 enough 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 a distributional semantics measure based on an algorithm provided by the increasingly popular word2vec tool. 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, it appears 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.

Keywords: Topic modeling, Topic coherence, LDA, NMF

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, and 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, and Mimno, 2009b). 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, and 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, and Baldwin, 2010; Mimno, Wallach, Talley, Leenders, and McCallum, 2011; Lau, Newman, and Baldwin, 2014) and distributional semantics (Aletras and 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, and Harshman, 1990) or Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999; Arora, Ge, and Moitra, 2012b). 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, and 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):

- 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

*Corresponding author

Email addresses: derek.o-callaghan@ucdconnect.ie (Derek O'Callaghan), derek.greene@ucd.ie (Derek Greene), joe.carthy@ucd.ie (Joe Carthy), padraig.cunningham@ucd.ie (Pádraig Cunningham)

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 appearance of terms in multiple descriptors for a particular model, signifying the presence 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 appear to be commonly used, with popular implementations being run accordingly (McCallum, 2002; Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011), in addition to recommended parameter values (Steyvers and Griffiths, 2006). Two out of three coherence measures, including a new measure based on word2vec (Mikolov, Chen, Corrado, and Dean, 2013a) 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, it appears 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 and Griffiths, 2006). This strategy dates back to the early work on latent semantic indexing by Deerwester, Dumais, Landauer, Furnas, and Harshman (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 and 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, Ng, and Jordan (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 Figure 1.

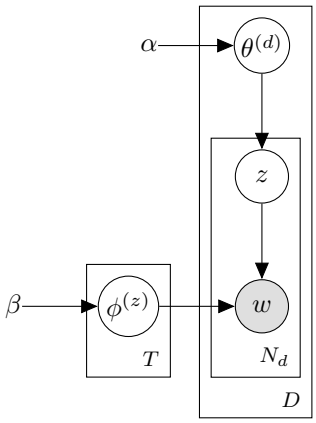


Figure 1: Plate notation for the graphical LDA topic model.

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:

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

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:

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

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 and 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, Wallach, Talley, Leenders, and McCallum (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 and Seung, 1999). Although it has been used in multiple domains, it is also applicable to topic modeling (Arora, Ge, and Moitra, 2012b). 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 and Seung, 2001) used to generate W and H are the Euclidean squared error:

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

and the Kullback-Leibler (KL) divergence, when V and WH both sum to 1 (thus acting as normalized probability distributions):

$$D(V||WH) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}}\right) \qquad (4)$$

NMF with KL divergence was previously shown to be equivalent to PLSA by Gaussier and Goutte (2005). As an alternative to the multiplicative update rules approach of Lee and Seung (2001) for determining W and H, Lin (2007) proposed the use of a projected gradient method with alternating non-negative least squares. Separately, to address the instability introduced by standard random initialization of W and H, Boutsidis and Gallopoulos (2008) introduced deterministic initialization with Non-Negative Double Singular Value Decomposition (NNDSVD), which is particularly suitable for sparse matrices. As with LDA, here we are primarily concerned with popular NMF variants, where the focus is upon the implementation of the method proposed by Lin (2007), with the squared error objective function and NNDSVD initialization, as provided by the *scikit-learn* machine learning library (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011).

2.2. Topic Coherence

Although perplexity (held-out likelihood) has been a common method for the evaluation of topic models, the study of Chang, Boyd-Graber, Gerrish, Wang, and Blei (2009) found that this was often negatively correlated with human judgements of topic quality (using topics discovered with PLSI, CTM, and LDA), and suggested that evaluation should be focused upon real-world task performance that includes human annotation. This has led to a number of studies that have focused upon the development of topic *coherence* measures, which capture the semantic interpretability of discovered topics based on their corresponding description terms (for example, the top N most probable terms from a ϕ distribution estimated by LDA). Newman, Lau, Grieser, and Baldwin (2010) calculated the correlation between human judgements and a set of proposed measures, and found that a Pointwise Mutual Information (PMI) measure achieved best or near-best out of all evaluated measures. This was based on co-occurrence frequency of each set of top 10 (LDA) topic terms within a reference corpus (Wikipedia), using a sliding window of 10 words, with the mean pairwise term PMI used as an individual topic score, where the intuition was that terms that regularly co-occurred were likely to produce coherent topic descriptors. A similar co-occurrence measure was suggested by Mimno, Wallach, Talley, Leenders, and McCallum (2011), which used log conditional probability (LCP) rather than PMI (conditioned on the higher-ranking term in each term pair), and was found to produce higher correlation with human judgements than that of the latter. In contrast to Newman, Lau, Grieser, and Baldwin (2010), the co-occurrence frequencies were calculated using the corpus being modeled, rather than relying upon a reference corpus. Here, LDA topics were discovered using the Gibbs sampling method along with their proposed extension.

Both of these measures were employed in the study of Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012), which compared the coherence of topics generated by LSA, NMF and LDA using a model-level summarization. Although they found that each of these methods had certain strengths, they also concluded that NMF tended to produce more incoherent topics than either of the other two methods. Aletras and Stevenson (2013) proposed measuring LDA topic coherence using distributional similarity (DS) between the top terms, where each term was represented as a vector in a semantic space, with topic coherence calculated as mean pairwise vector similarity; Cosine sim-

ilarity, Jaccard similarity, and the Dice coefficient were used. As before, these were correlated with human judgements, in addition to the PMI measure, a normalized variant of PMI (Bouma, 2009) (NPMI, range = $[-1, 1]$), and LCP. They found the term vector similarity measures to compare favorably with those based on PMI, in particular, Cosine similarity, while LCP performed poorly in general, where they suggested that the latter is sensitive to the size of the modeled corpus. Lau, Newman, and Baldwin (2014) performed an empirical comparison of these four PMI, NPMI, LCP and DS measures in the context of the original evaluation tasks used by Chang, Boyd-Graber, Gerrish, Wang, and Blei (2009) (using PLSI, LDA, and CTM), where the NPMI and DS measures were those most strongly correlated with the human raters at the topic level. To address the sparsity issue with using the corpus being modeled to calculate term co-occurrence for LCP, as pointed out by Aletras and Stevenson (2013) (and also an earlier work by Lau, Baldwin, and Newman, 2013), they proposed instead using the same reference corpus as required by the PMI measures.

More recently, Röder, Both, and Hinneburg (2015) proposed a unifying framework that represented coherence measures as a composition of parts, with the objective of achieving higher correlation with human judgements. This was an attempt to address certain issues raised in their earlier work in relation to coherence measures based on term co-occurrence (Rosner, Hinneburg, Röder, Nettling, and Both, 2013). The emphasis was largely on topics discovered by LDA, as they followed the evaluation schemes of Newman, Lau, Grieser, and Baldwin (2010), Aletras and Stevenson (2013), and Lau, Newman, and Baldwin (2014). One clear observation that can be made about these previous works is the attention given to LDA or similar probabilistic topic modeling methods, where matrix factorization methods such as NMF are very much in the minority. The current work aims to address this issue with a comparison of topics discovered with NMF and LDA across multiple corpora, particularly in light of our previous findings that the former produces more readily-interpretable topics (O'Callaghan, Greene, Conway, Carthy, and Cunningham, 2013).

3. Data

A range of corpora were analyzed in this evaluation, where we were focused upon both new and existing corpora containing documents that had been (manually) annotated with classes. The first new corpus contained news articles from the BBC website[1]. At the start of January 2014, we retrieved all tweets up to the 3,200 Twitter REST API[2] limit for 71 manually selected Twitter accounts affiliated with the BBC or its journalists (for example, @BBCNews, @BBCSport, @bbcscitech), which yielded a total of 91,616 tweets. All unique URIs containing the domains *bbc.co.uk* or *bbc.com* were extracted from these tweets, and the corresponding web pages (where still accessible) were then retrieved, with the exception of a set of blacklisted URI prefixes (for example, those related to careers, advertising and other non-news content). From the retrieved pages, we extracted and fetched web pages for all unique URIs containing either of the two BBC domains that had not been previously fetched; this process was performed twice. We then filtered all retrieved articles that were published outside the time period 2009-01-01 to 2013-12-31, or whose publication date could not be ascertained from the corresponding page metadata. The article body text was extracted using the Java Boilerpipe library[3] (Kohlschütter, Fankhauser, and Nejdl, 2010), with articles containing empty body text being filtered. The final corpus consisted of all articles in the top 40 *sections*, as annotated by the BBC and extracted from the page metadata, where each article is assigned to exactly one section.

A similar process was used to generate a corpus containing news articles from the Guardian website[4], where 29 manually selected Twitter accounts affiliated with the Guardian or its journalists (for example, @guardian, @GdnPolitics, @guardianfilm) yielded a total of 184,284 tweets. Unique URIs containing the domains *guardian.co.uk* or *theguardian.com* were used to fetch the article text, with the final corpus consisting of articles found in the top 24 annotated *sections* as extracted from the metadata, with each article featuring one section annotation. Two additional news corpora were extracted from the New York Times Corpus, which contains over 1.8 million articles written and published by the New York Times (NYT) between January 1, 1987 and June 19, 2007, in addition to article metadata annotations (Sandhaus, 2008). The first of these corpora consisted of articles from 2003, as also analyzed by Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012) in their coherence evaluation of multiple topic modeling methods, with the second containing a stratified 10% sample of articles from 2000 to 2007. In contrast to our extracted BBC and Guardian corpora, the NYT articles can contain multiple section annotations. For both extracted corpora, identifiable *meta*-sections covering multiple topics were excluded, such as "Front Page", "Corrections", and "Magazine".

The final two corpora consisted of pages found on Wikipedia. We initially selected eight top-level categories from the WikiProject Council Directory[5] that had active sub-categories and/or task forces, and also selected the set of active sub-categories. As the sub-categories found in each top-level category directory are often not peers of each other in the hierarchy, sub-categories at the highest level were chosen, for example a sport category was selected as opposed to a category for a particular team. For each category, a list of page titles in the classes 'FA', 'FL', 'A', 'GA', 'B', 'C' was retrieved from the corresponding category class page, and the page text was extracted from a Wikipedia dump from January 2014 by means of a wrapper around the WikiExtractor utility[6]. This resulted in 42,170 page titles for

[1] http://www.bbc.com

[2] https://dev.twitter.com/docs/api

[3] https://code.google.com/p/boilerpipe/

[4] http://www.theguardian.com

[5] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Directory

[6] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

	Documents	Terms	Classes
BBC	161,469	17,079	40
Guardian	194,153	22,141	24
NYT 2003	70,134	20,429	20
NYT 2000+ (10%)	65,335	21,461	20
Wikipedia (high-level)	5,682	28,699	6
Wikipedia (lower-level)	4,970	24,265	10

(a) Corpus size

	Min	Max	Mean	Median	Standard deviation
BBC	10	7241	242.00	167	362.53
Guardian	10	19080	299.08	250	276.65
NYT 2003	10	7816	282.40	260	239.42
NYT 2000+ (10%)	10	16997	302.88	275	298.26
Wikipedia (high-level)	10	7510	897.35	607	901.03
Wikipedia (lower-level)	10	7520	924.31	604	929.62

(b) Document length statistics (terms)

Table 1: The six corpora used in the topic coherence evaluation, including the number of documents and terms following pre-processing (described in Section 4.1).

194 categories containing text with ≥ 10 terms, with some titles belonging to multiple categories. The two corpora consisted of a selection of top-level categories and lower-level subcategories respectively. Details of all six corpora used in the evaluation can be found in Table 1, and pre-processed versions are made available online for further research[7].

4. Methodology

4.1. Topic Discovery and Descriptor Generation

A set of common pre-processing steps were applied to all six corpora. We initially compiled a "standard" set of 671 English stopwords from a variety of sources including those featured in the machine learning library scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011), the Natural Language Toolkit (NLTK) (Loper and Bird, 2002), and the MALLET toolkit (McCallum, 2002), in addition to English honorifics and contractions documented on Wikipedia[8] [9]. These stopwords were filtered, along with any terms containing common top-level-domains such as ".com" or ".co.uk". URIs and terms containing digits were also filtered, and ascii normalization was performed to remove diacritics. Terms were converted to lowercase, and a lemmatizer was applied (NLTK wrapper around WordNet's built-in morphy function). Stemming was not performed as it often led to the subsequent generation of topic descriptor terms that were not interpretable by an end user. Finally, low-frequency terms occurring in fewer than l of the total m documents were also excluded, where the l threshold was set to $\max(10, m/1000)$.

Although multiple variants exist for both LDA and NMF (for example, Blei, Griffiths, Jordan, and Tenenbaum, 2004; Rosen-Zvi, Griffiths, Steyvers, and Smyth, 2004; Saha and Sindhwani, 2012), we restricted the experiments to those that are popular and/or were used by the topic coherence papers discussed earlier in Section 2.1. In the case of LDA, we used the fast Gibbs sampling implementation provided by the MALLET toolkit (McCallum, 2002), with the same parameters as found in the coherence evaluation code provided by Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012) [10] apart from an increased number of iterations. This included the recommended values (Steyvers and Griffiths, 2006) for the Dirichlet hyperparameters α and β of $50/k$ (k = number of topics) and 0.01 respectively (these are actually the default parameter values in MALLET), in addition to hyperparameter optimization for an asymmetric Dirichlet prior over the document-topic distribution θ (Wallach, Mimno, and McCallum, 2009a). LDA operates on *bag-of-words* document representations, and the corresponding feature sequences used by MALLET were created for each corpus following the pre-processing steps described above.

For NMF, the same pre-processed corpus documents were transformed to log-based Term Frequency-Inverse Document Frequency (TF-IDF) vectors (Salton and Buckley, 1988), and subsequently normalized to unit length. We used the squared error NMF variant as provided by scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011), which is an implementation of the fast alternating least squares method proposed by Lin (2007). The Kullback-Leibler objective function was not considered for this evaluation due to its equivalence to PLSA (Gaussier and Goutte, 2005), itself a probabilistic precursor to LDA. To address the instability introduced by random initialization in standard NMF, the deterministic NNDSVD initialization method was also employed (Boutsidis and Gallopoulos, 2008).

For each topic found by applying these variants of LDA and NMF, a descriptor was created as follows:

1. NMF^w: the top 10 highest-ranking terms from the topic's basis vector in W^k, a factor of V_{TF-IDF}.

2. LDA^u: the top 10 most probable terms from the topic's ϕ distribution.

The w (weighted) and u (unweighted) notation reflects the term weighting strategy employed, in addition to the simplest *bag-of-words* weighting strategy based on term frequencies. The IDF pre-processing step used by NMF^w down-weights the contribution of TF in the case of frequent (more general) terms, while also boosting the contribution of rarer terms that may be more discriminating. As the analogous term weighting with LDA^u is effectively that of TF, we refer to it as unweighted for the purpose of this evaluation. Although it is customary to generate LDA term descriptors using the most probable terms, Blei and Lafferty (2009) state their preference for selecting the top terms ranked using the score defined in Equation 5:

$$\text{term-score}_{k,v} = \widehat{\beta}_{k,v} \log \left(\frac{\widehat{\beta}_{k,v}}{(\prod_{j=1}^{K} \widehat{\beta}_{j,v})^{\frac{1}{K}}} \right) \qquad (5)$$

[7] http://mlg.ucd.ie/topiccoherence/

[8] http://en.wikipedia.org/wiki/English_honorifics

[9] http://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

[10] https://github.com/fozziethebeat/TopicModelComparison

This is inspired by TF-IDF weighting, where the first expression $\beta_{k,v}$, the probability of term v for topic k, is analogous to TF, while the second expression down-weights terms that have high probability across all k topics, somewhat similar to IDF. As this operation mirrors the weighted nature of NMF^w, we also discovered topics by applying NMF to TF input vectors V_{TF} (minus the IDF component), which in turn mirrors the unweighted nature of LDA^u. Thus, the following topic descriptors are also generated:

3. NMF^u: the top 10 highest-ranking terms from the topic's basis vector in W^k, a factor of V_{TF}. As this is based on a pre-processing operation, these topics are entirely separate to those of NMF^w.

4. LDA^w: the top 10 highest-ranking terms, weighted using Equation 5. As this is based on a post-processing operation, these topics are the same as those used by LDA^u

At this point, we note that other work that measured coherence of topics found by LDA appears to be largely focused upon the LDA^u topic term descriptors of option 2 (Newman, Lau, Grieser, and Baldwin, 2010; Mimno, Wallach, Talley, Leenders, and McCallum, 2011; Stevens, Kegelmeyer, Andrzejewski, and Buttler, 2012; Aletras and Stevenson, 2013; Lau, Newman, and Baldwin, 2014).

4.2. Measuring Topic Coherence and Generality

Having generated a set of topic models, the following topic coherence measures were calculated for each of the four term descriptor methods NMF^w, LDA^u, NMF^u, and LDA^w, with $N = 10$:

1. TC-NPMI - Normalized PMI (Aletras and Stevenson, 2013; Lau, Newman, and Baldwin, 2014) (the unnormalized version originally proposed by Newman, Lau, Grieser, and Baldwin (2010) was also calculated, but the normalized version is reported here given its superior performance as demonstrated by Lau, Newman, and Baldwin, 2014):

$$\text{TC-NPMI} = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P(w_j,w_i)+\epsilon}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)+\epsilon} \quad (6)$$

2. TC-LCP - Mean pairwise log conditional probability (Mimno, Wallach, Talley, Leenders, and McCallum, 2011):

$$\text{TC-LCP} = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{P(w_j,w_i)+\epsilon}{P(w_i)} \quad (7)$$

3. TC-W2V - As an analog to the DS measures of Aletras and Stevenson (2013), we propose the creation of term vectors wv using a **word2vec** model (Mikolov, Chen, Corrado, and Dean, 2013a). This tool provides two neural network-based algorithms for estimating word representations in a vector space; Continuous Bag-of-Words (CBOW), where the current word is predicted based on its context, and Skip-gram, which predicts context words based on the current word. These approaches have been found to generate word vectors that explicitly encode linguistic regularities from large amounts of unstructured text data (Mikolov, Sutskever, Chen, Corrado, and Dean, 2013b), and so are appropriate for use with a large reference corpus in the analysis of topic coherence. Here, the coherence score is the mean pairwise Cosine similarity of two term vectors generated with a Skip-gram model:

$$\text{TC-W2V} = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} similarity(wv_j, wv_i) \quad (8)$$

For each coherence measure, we generated an aggregate score for a particular (*descriptor method*, k) model by taking the mean of the constituent topic scores. Similar model-level coherence scores were also used in the evaluation of Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012). As suggested by Mimno, Wallach, Talley, Leenders, and McCallum (2011), a smoothing count $\epsilon = 1$ was included as required to avoid taking the logarithm of zero.

In addition to measuring topic coherence, we also analyzed the generality of the topic descriptors for the four methods NMF^w, LDA^u, NMF^u, and LDA^w. Here, generality is considered at two levels; 1) the overlap or dependence between topics, based on the appearance of terms in multiple descriptors for a particular model, and 2) the tendency for a method to generate topics containing high-frequency descriptor terms from the underlying corpus. As discussed by Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, and Zhu (2012a), some level of similarity is to be expected, but lower numbers of unique terms across topic descriptors can be an indication of less useful models. The following steps were performed for each (*descriptor method*, k) model:

1. The mean pairwise Jaccard similarity between the topic descriptors TD was calculated. Higher similarity values indicate increased topic dependency:

$$\text{MPJ}_{m,k} = \frac{1}{\binom{k}{2}} \sum_{j=2}^{k} \sum_{i=1}^{j-1} \frac{|TD_i \cap TD_j|}{|TD_i \cup TD_j|} \quad (9)$$

2. The probability distribution of descriptor term occurrences across all topics was generated. Terms having high occurrence frequencies often appear to be general terms from the underlying corpus.

Only the top N (10) descriptor terms are considered, as these are the terms that would be presented to an end user. The generality of the remaining terms for a particular topic is effectively irrelevant. In addition, we do not consider the raw term values due to the different methods being evaluated.

5. Evaluation

In this section, we provide an evaluation of the four topic descriptor methods NMF^w, NMF^u, LDA^u, and LDA^w in terms

of their corresponding coherence and generality, using the model-level measures described in Section 4. We also analyze the associated document-topic memberships by referencing the underlying corpus class labels. As we find similar patterns across all six corpora, we illustrate the differences in coherence and generality between the methods by focusing specifically on the BBC corpus results.

5.1. Model coherence

Tokenized versions of the documents belonging to each of the six corpora were generated using the pre-processing steps described in Section 4.1, where the final statistics can be found in Table 1. For each corpus, the documents were transformed to log-based TF-IDF unit vectors, topics were discovered using the scikit-learn implementation of NMF (including NNDSVD initialization), and the corresponding NMF^w topic descriptors were generated from the highest-ranking top 10 terms found in each topic basis vector. Similarly, NMF was also applied to TF vector representations of the documents to generate the corresponding NMF^u topic descriptors. In the case of LDA, the MALLET implementation was applied to the sets of document feature sequences, from which the LDA^u and LDA^w topic descriptors were generated respectively from the top 10 most probable topic terms, and the top 10 topic terms following the Blei and Lafferty normalization described in Equation 5. In all cases, topics were discovered for values of $k \in [10, 100]$ (intervals of 10), where this seemed to be a reasonable range given the number of annotated classes (see Table 1), using the parameters for NMF and LDA as described in Section 4.1.

Coherence scores were then calculated for the four topic descriptor methods, with co-occurrence frequencies generated for the unique descriptor terms across all models using a reference corpus for TC-NPMI and also for TC-LCP, due to the issues associated with using the corpus being modeled for the latter that were highlighted by Aletras and Stevenson (2013), and Lau, Newman, and Baldwin (2014). This reference corpus consisted of a Wikipedia dump from January 2014, where the tokenization process included filtering the same 671 stopwords as used in pre-processing of the six corpora used for evaluation, term lemmatization, with all remaining terms retained. Following tokenization, the term co-occurrence frequencies were calculated using a sliding window of 20 terms. This tokenized Wikipedia corpus was also used to create the word2vec model as required by TC-W2V, using the same parameters as the *demo-word.sh* script provided with revision 37 of the source code[11], i.e. the Skip-gram model with word vector dimensions = 200, max context window skip length = 5, hierarchical softmax enabled, negative sampling disabled, and sample threshold = $1e$-3. The word2vec Cosine similarity between each pair of unique topic descriptor terms was calculated at this point.

Mean model-level coherence scores for the four topic descriptor methods are presented for TC-NPMI, TC-LCP and TC-W2V in Figure 2, Figure 3, and Figure 4 respectively. The coherence score scale is less important here, where the relative difference between the methods is more interesting. For TC-NPMI and TC-W2V, a certain level of separation is observable between the weighted (w) and unweighted (u) topic descriptor methods. It appears that the weighted methods are producing more coherent topics, where NMF^w is regularly the most coherent method with LDA^w also performing strongly, while the model-level coherence of the LDA^u topic descriptors (generated from the most probable terms for a particular topic) is always lower. This pattern is replicated across all six corpora. However, the situation seems to be somewhat reversed in the case of TC-LCP, where LDA^u is found to be most coherent, with the NMF methods performing poorly and LDA^w positioned in-between.

We now illustrate the differences between TC-LCP and the other two measures with an analysis of the coherence scores of two pairs of terms that were included in one of the BBC topic descriptors that appears to education-related, where these scores can be found in Table 2. The term pair *(school, year)* yields a higher TC-LCP score than that of *(school, education)*, while the reverse is true for the other measures. At a glance,

	w_i = school, w_j = year	w_i = school, w_j = education
$\boldsymbol{P(w_j)}$	0.08	0.01
$\boldsymbol{P(w_i, w_j)}$	0.005	0.004
TC-NPMI	0.12	0.38
TC-LCP	-0.85	-1.01
TC-W2V	0.20	0.54

Table 2: Examples of pairwise differences between TC-LCP and the other three coherence measures. TC-LCP tends to produce higher scores when one or both terms are more general, this can be seen with the score from the pair containing the general term $year$ ($P(w_j) = 0.08$) compared to that of the other pair with the more specific term $education$ ($P(w_j) = 0.01$).

Term	Probability
manager	0.06
uk	0.05
nation	0.05
final	0.05
country	0.05
year	0.04
win	0.04
team	0.04
market	0.04
championship	0.04

(a) NMF^w

Term	Probability
league	0.13
involved	0.10
ball	0.09
team	0.08
wicket	0.07
shot	0.06
point	0.06
world	0.05
minute	0.05
match	0.05

(b) NMF^u

Term	Probability
year	**0.30**
people	**0.18**
goal	0.12
team	0.11
world	0.10
uk	0.09
league	0.08
bbc	0.08
england	0.07
match	0.06

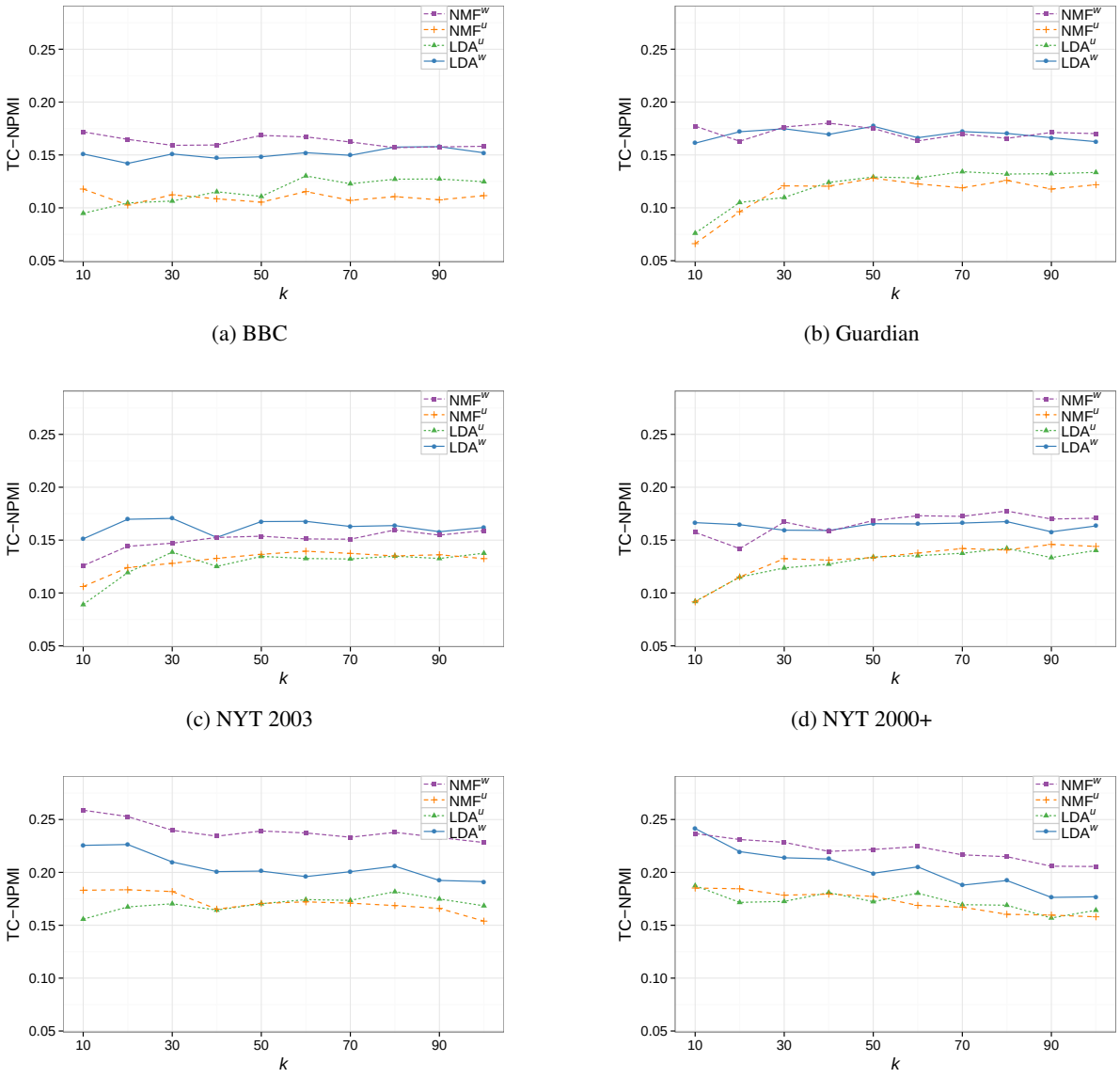
(c) LDA^u

Term	Probability
people	**0.11**
league	0.08
goal	0.08
england	0.07
world	0.06
team	0.06
player	0.06
ball	0.06
season	0.05
match	0.05

(d) LDA^w

Table 3: Top ten most frequent topic descriptor terms for the BBC corpus with $k = 100$.

[11] https://code.google.com/p/word2vec/



(a) BBC

(b) Guardian

(c) NYT 2003

(d) NYT 2000+

(e) Wikipedia (high-level)

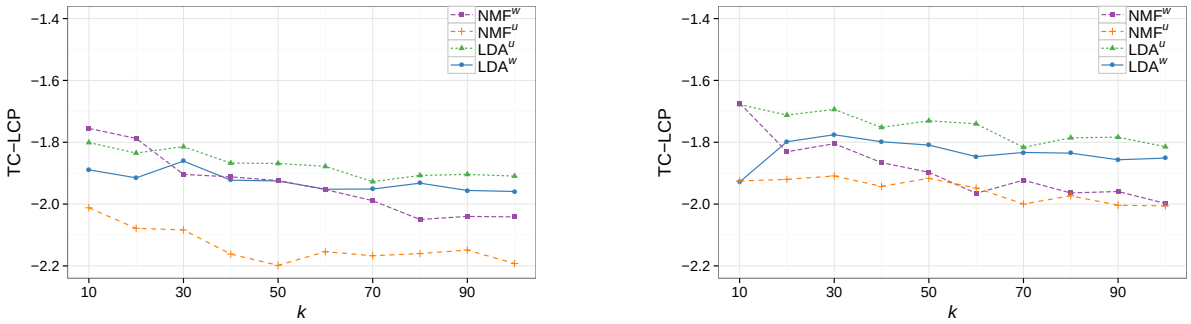
(f) Wikipedia (lower-level)

Figure 2: Mean topic TC-NPMI coherence scores for $k \in [10, 100]$.

both term pairs appear coherent. However, it might be argued that *(school, education)* is somewhat more coherent than *(school, year)*, due to the general nature of the *year* term. TC-NPMI, being based on PMI, considers the probability of both terms in a particular term pair, where frequent terms will be down-weighted with respect to those that occur less frequently, such as *education* from this example. Lau, Newman, and Baldwin (2014) discuss the bias of the unnormalized version of this measure (TC-PMI) towards lower frequent terms, which should be corrected by TC-NPMI. At the same time, this example clearly demonstrates that the TC-NPMI score is higher for the *(school, education)* pair, which is further supported by the corresponding word2vec Cosine similarity. As the TC-LCP measure only considers the probability of one (the highest-ranking) term for a particular term pair, the appearance of general terms is less of an issue, particularly when both terms are general. This behaviour, coupled with the tendency for LDA to generate high-ranking topic terms that are more general (Chemudugunta, Smyth, and Steyvers, 2006) is likely the reason for the higher LDA^u TC-LCP scores.

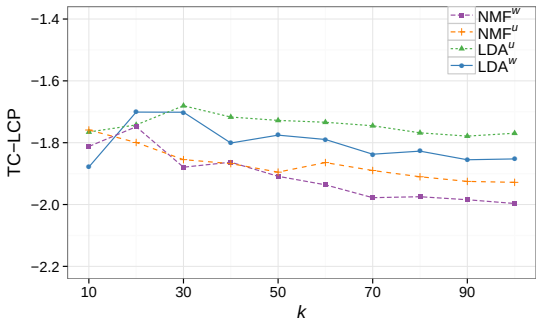
5.2. *Model Generality*

We also analyzed the generality of topic descriptors produced by the NMF^w, NMF^u, LDA^u, and LDA^w methods, where we were specifically interested in the overlap or dependence between topics, based on the appearance of terms in multiple descriptors for a particular model, and the tendency for a method to generate topics containing high-frequency descriptor terms from the underlying corpus. The mean Jaccard similarity between the topic descriptors generated by all four methods was

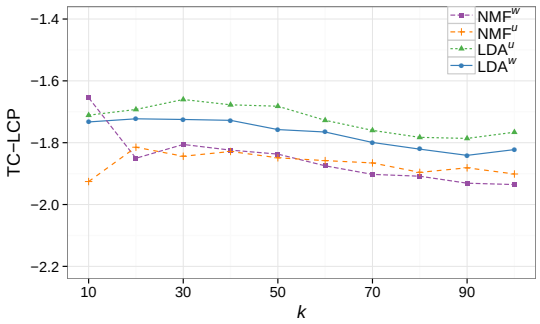


(a) BBC

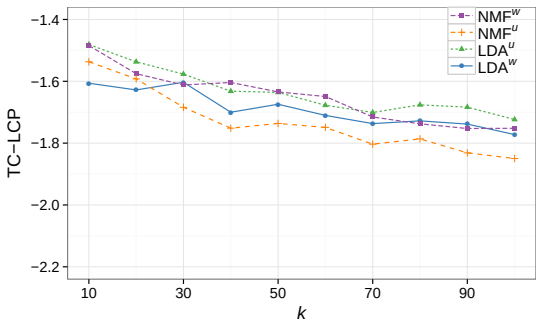
(b) Guardian



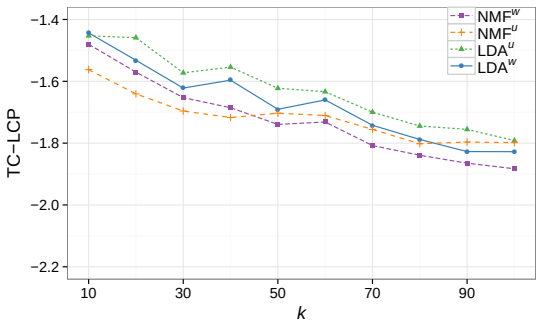
(c) NYT 2003



(d) NYT 2000+



(e) Wikipedia (high-level)

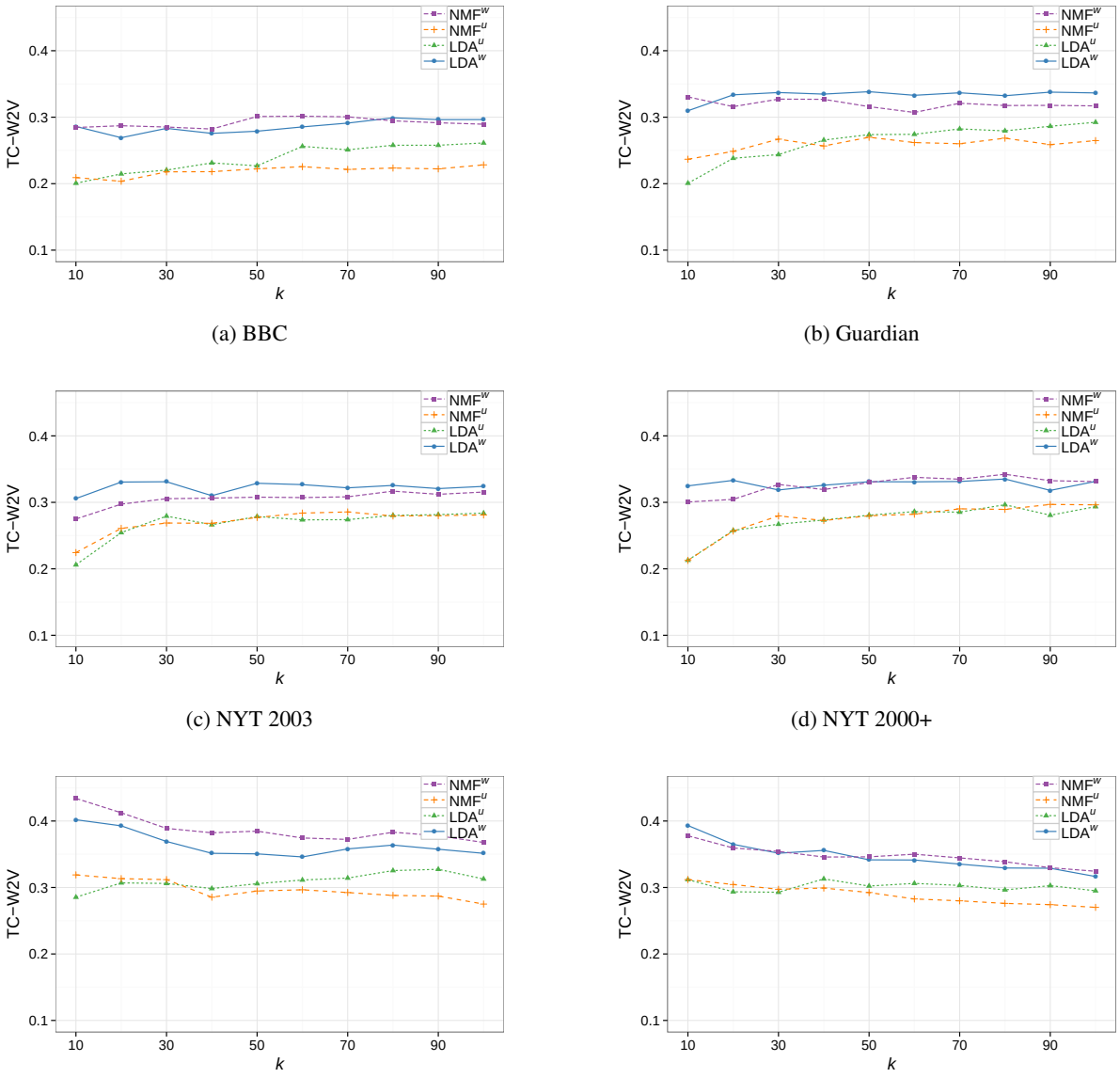


(f) Wikpedia (lower-level)

Figure 3: Mean topic TC-LCP coherence scores for $k \in [10, 100]$.

calculated, and the results for values of $k \in [10, 100]$ are presented in Figure 5. As with the coherence scores discussed in Section 5.1, a pattern is observable across all six corpora, where the highest levels of similarity are always found with the LDA^u descriptors, with NMF^w producing those that are least similar in most cases. LDA^w also generates relatively low levels of similarity, while it is interesting to note that there is a separation between the unweighted methods LDA^u and NMF^u, where the similarity of the latter is closer to those of the weighted methods. The overlap in topics produced by LDA^u due to lower numbers of unique terms across its topic descriptors may be an indication of less useful (coherent) models, as suggested by Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, and Zhu (2012a). It is likely that the different behaviour observed with the Wikipedia corpora is related to the corresponding smaller number of annotated classes. Here, the pattern at $k = 10$ (the value of k that is closest to the actual number of classes) appears similar to that of the other corpora.

This overlap can also be demonstrated by looking at the frequency of terms occurring in X (multiple) descriptors, where Figure 6 contains the results for $X \in [2, 3, 4, \geq 5]$ from all four methods, with $k = 100$. Here, it can be seen that LDA^u consistently generates higher frequencies of terms occurring in $X \geq 5$ descriptors. Further investigation finds that these are often general terms from the underlying corpus. For example, Table 3 contains the top ten most frequent topic descriptor terms for the BBC corpus with $k = 100$. Although certain terms are occurring frequently for all four methods (also seen in Figure 6), those of LDA^u appear to be general terms that may be less discriminating when coherence is considered, with *year* and *peo-*



(a) BBC

(b) Guardian

(c) NYT 2003

(d) NYT 2000+

(e) Wikipedia (high-level)

(f) Wikpedia (lower-level)

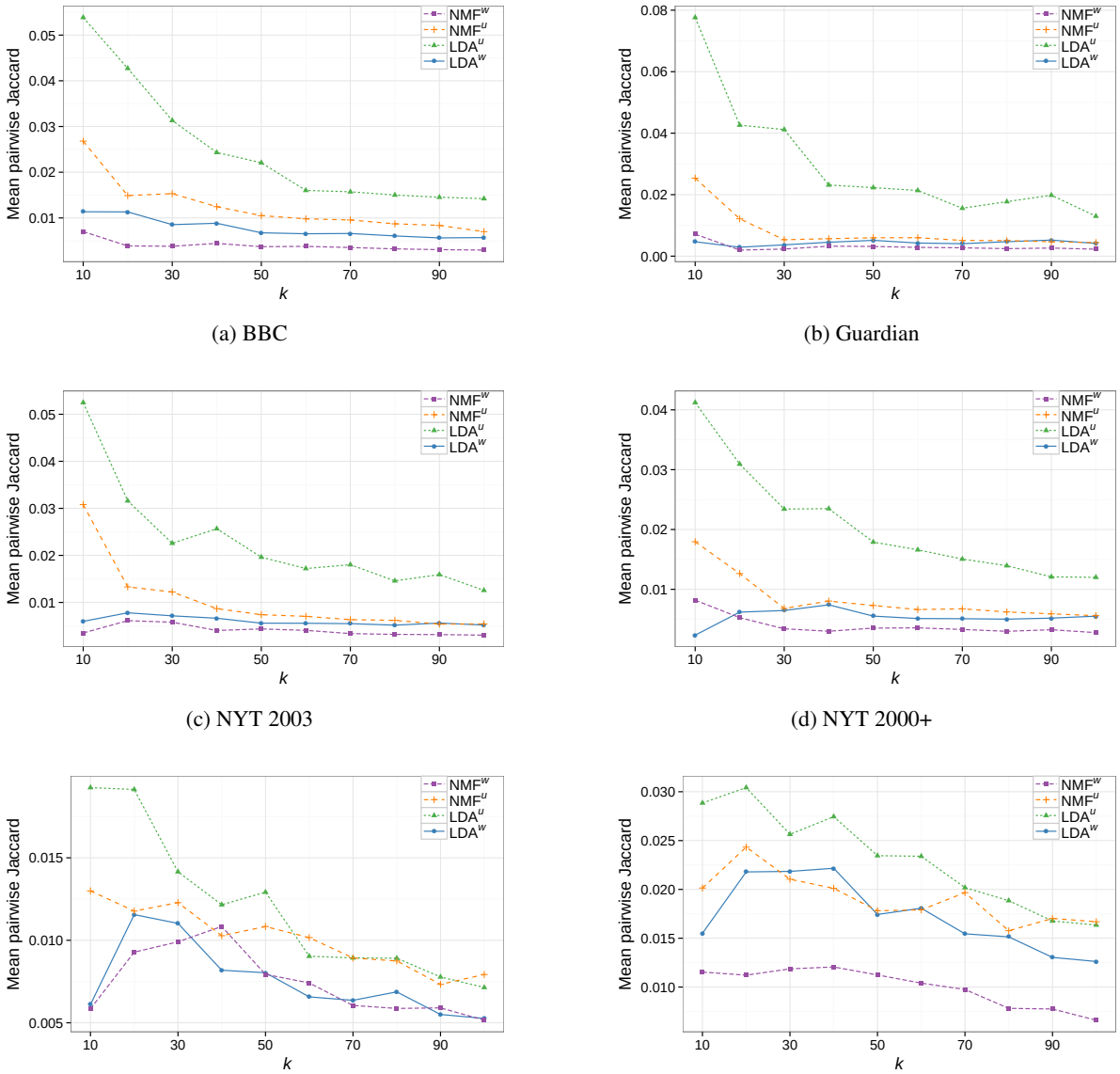
Figure 4: Mean topic TC-W2V coherence scores for $k \in [10, 100]$.

ple occurring in 30% and 18% of topic descriptors respectively. In the case of LDA^w, *people* is also highly-ranked, albeit with lower probability. Both it and NMF^u feature football-related terms, indicating the presence of topic overlap. A number of general terms such as *uk* and *year* are also present with NMF^w. However, their relatively lower probability suggests that this is less of an issue in comparison to the other methods.

5.3. Document-Topic Membership

Our main objectives in this work were the analysis of topic descriptor coherence and generality. In addition, we completed our evaluation with a brief look at the agreement between the document-topic membership and the underlying corpus class labels. Although both LDA and NMF permit the assignment of documents to multiple topics, we focused solely on disjoint analysis where membership was assigned using the highest-ranking topic for each document. This can be justified due to the fact that the classes in the six corpora are also largely disjoint. All documents assigned to multiple classes were excluded. We used Adjusted Mutual Information (AMI) to measure this agreement, which produces results in the range $[0, 1]$ and corrects for chance agreement, while also accounting for the fact that MI tends to be larger with clusterings containing higher numbers of clusters.

AMI scores for NMF^w, NMF^u, and LDA^u ($k \in [10, 100]$) can be found in Figure 7. The LDA^w topic descriptor method is not included here as its descriptors are derived from the post-processed LDA topic-term distributions; it has the same document-topic distributions as LDA^u. The agreement scores are relatively low for the non-Wikipedia corpora, where LDA^u pro-



(a) BBC

(b) Guardian

(c) NYT 2003

(d) NYT 2000+

(e) Wikipedia (higher-level)

(f) Wikipedia (lower-level)

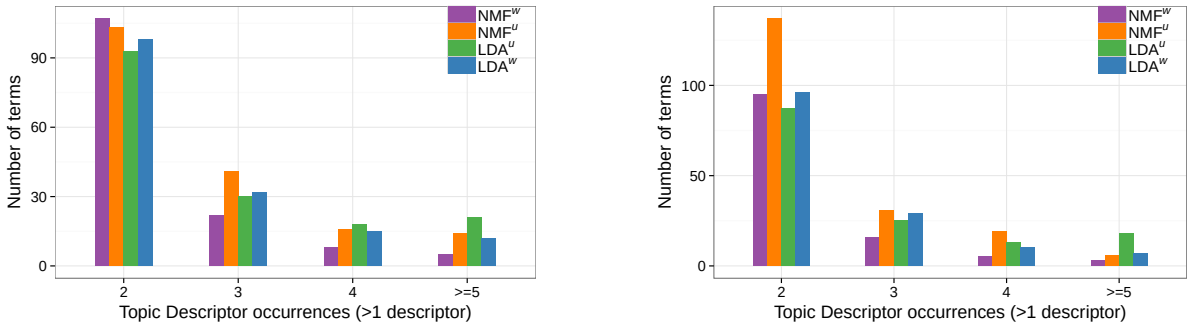
Figure 5: Mean pairwise Jaccard similarity of topic descriptors (using top 10 topic terms) for $k \in [10, 100]$.

duces slightly higher scores than NMF^w, with NMF^u performing poorly in all cases. Higher agreement scores with little difference between the methods are observed with Wikipedia. It is likely that these results are related to the smaller number of annotated classes in the Wikipedia corpora. They may also suggest the presence of a certain level of inaccurate document annotations.

5.4. Discussion

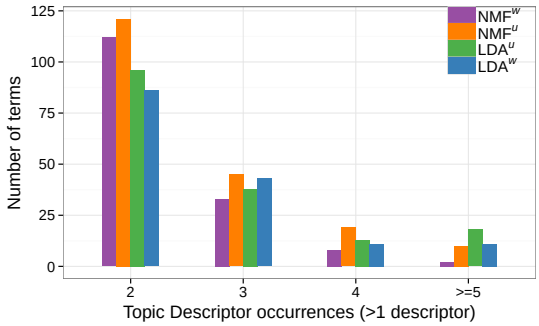
For all six corpora, we have observed differences between the scores generated by the weighted (NMF^w, LDA^w) and unweighted (LDA^u, NMF^u) topic descriptor methods. In the case of the aggregate model-level coherence scores, the weighted methods perform strongly for all measures apart from TC-LCP. This appears to contrast with the evaluation of Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012), where they found that the TC-LCP scores were often in agreement with those of the unnormalized version of TC-NPMI (the two coherence measures used in their evaluation), with the TC-LCP scores for NMF matching or exceeding those of LDA for $k \leq 100$. Lau, Newman, and Baldwin (2014) also observed strong correlation between the human coherence ratings and those of TC-LCP and TC-NPMI. However, as they did not compare topics discovered by the multiple methods they used with each other, it is unclear whether we can draw many parallels between their correlation-based findings and those of our own evaluation.

The measures related to topic generality find higher levels of similarity between the descriptors generated by LDA^u, along with the promotion of general high-frequency corpus terms among multiple descriptors. This effect is less noticeable with LDA^w,

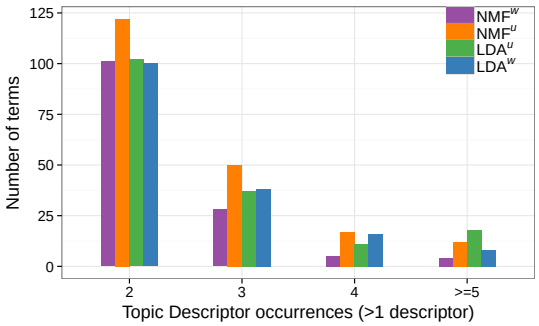


(a) BBC

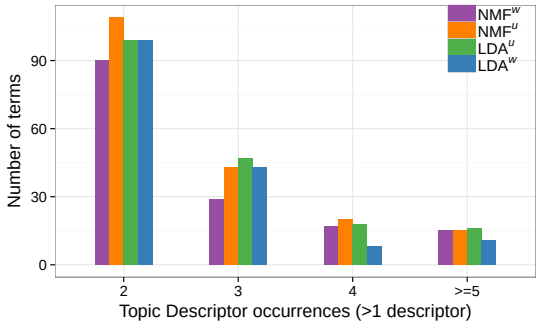
(b) Guardian



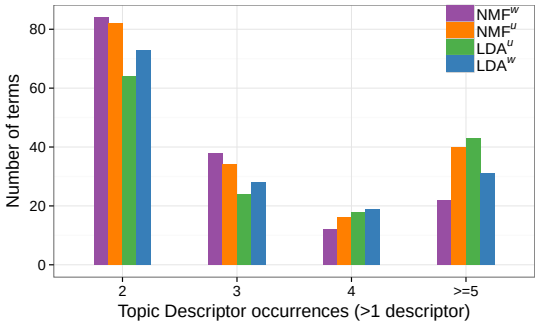
(c) NYT 2003



(d) NYT 2000+



(e) Wikipedia (higher-level)

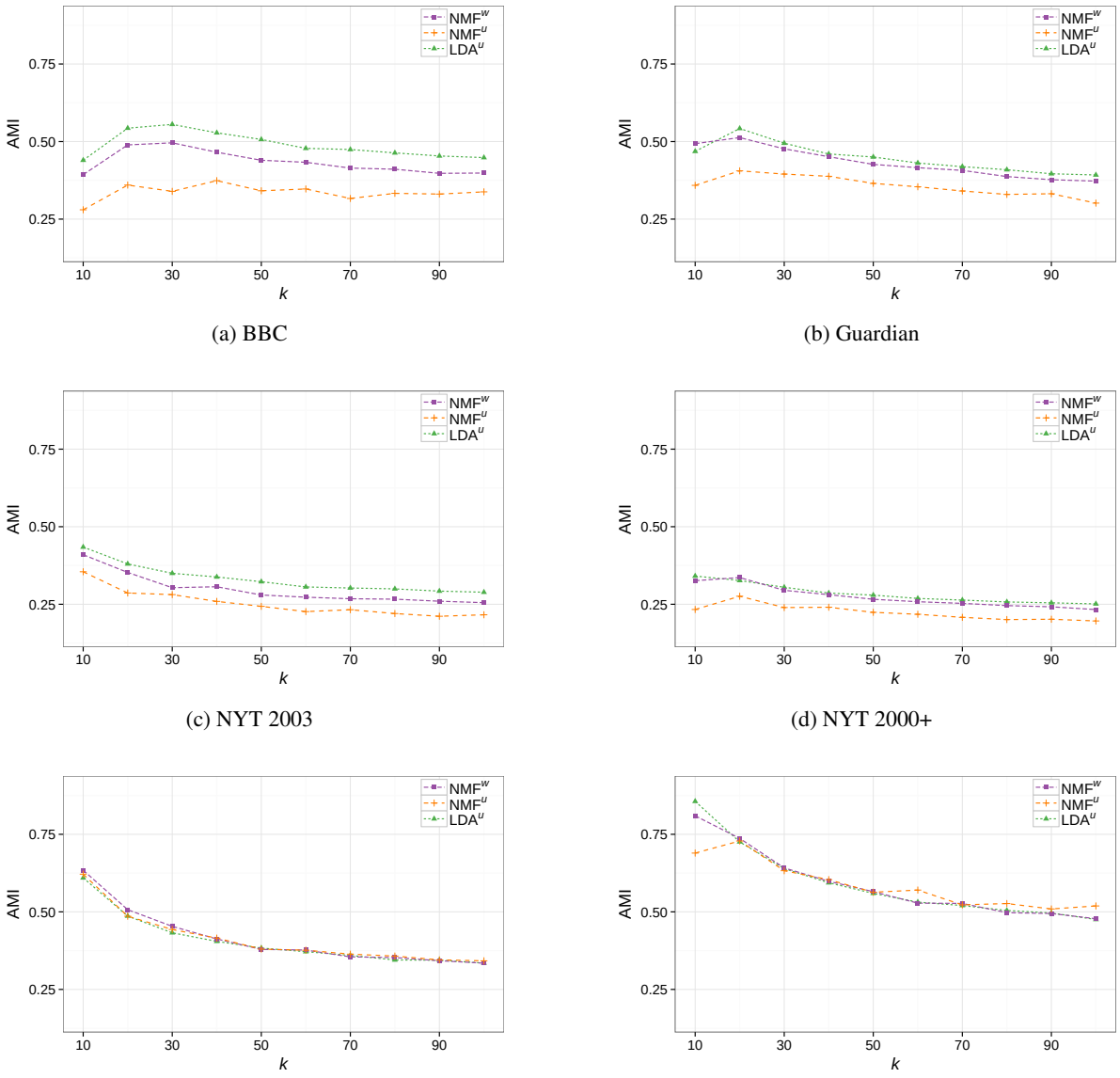


(f) Wikipedia (lower-level)

Figure 6: Frequency of terms occurring in X (multiple) topic descriptors, for $X \in [2, 3, 4, \geq 5]$ from all four methods with $k = 100$. LDA^u consistently generates higher frequencies of terms occurring in $X \geq 5$ descriptors.

and is significantly lower with NMF^w. Wallach, Mimno, and McCallum (2009a) have pointed out that it is often customary to create a corpus-specific stopword list to address this issue, even if some of these play meaningful semantic roles. They also suggest that using an asymmetric prior over the LDA document-topic distribution θ can result in topics that are unaffected by stopwords, with stopwords themselves being isolated in a small number of topics. However, although we have enabled this particular option in MALLET, we still observe the presence of general terms in multiple descriptors, which is only decreased when the Blei and Lafferty (2009) normalization of LDA^w is applied. We also note the differences in generality scores when the value of k is considerably different to the number of underlying corpus classes, as observed for both Wikipedia corpora with $k > 10$.

Steyvers and Griffiths (2006) suggested the use of probability distribution divergence measures such as Jensen-Shannon divergence when calculating the similarity between terms or documents following topic discovery; this is also applicable to the topics themselves. Although such measures can be applied to both LDA topics and those discovered by NMF, they were not employed here as we were specifically concerned with the top N topic descriptors that could ultimately be presented to an end user. Separately, Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012) also evaluated the impact of different ϵ values on the calculation of the two coherence measures they employed. They found that using a small value of $\epsilon = 10^{-12}$ resulted in a decrease in coherence scores for NMF compared to LDA, par-



(a) BBC

(b) Guardian

(c) NYT 2003

(d) NYT 2000+

(e) Wikipedia (higher-level)

(f) Wikipedia (lower-level)

Figure 7: AMI scores for NMF^w, NMF^u, and LDA^u, for $k \in [10, 100]$. LDA^w is not included as it is derived from the post-processed LDA topic-term distributions (it has the same document-topic distributions as LDA^u).

ticularly in the case of the PMI-based measure. We also calculated the TC-NPMI and TC-LCP scores using this ϵ value, where a similar decrease in coherence was observed for a subset of the NMF models. However, a close inspection of some of the coherence score distributions found that this small ϵ value produced significant outliers for term pairs that did not occur together in the reference Wikipedia corpus. The fact that an individual topic descriptor's score was calculated using the mean of the constituent term pairwise scores meant that it was sensitive to such outliers, which could lead to a low score even in the case of a single term pair not occurring together while the remaining term pairs were highly coherent. It might in fact be argued that taking the median of these pairwise scores is more appropriate. However, we felt that the presence of descriptor terms that do not occur together must be acknowledged, where the use of $\epsilon = 1$, as originally suggested by Mimno, Wallach, Talley, Leenders, and McCallum (2011), acts as a compromise between both extreme cases.

6. Conclusions

In this work, we have described an analysis of the semantic interpretability, also known as topic coherence, of the sets of top terms generally used to describe topics discovered by a particular algorithm. This has been achieved with an evaluation of popular variants of both probabilistic (LDA) and matrix decomposition (NMF) topic modeling techniques on multiple corpora, using a combination of existing and new measures that focus

on topic coherence and generality. A common pre-processing procedure has been employed for both techniques where possible, without relying on particular actions such as the generation of corpus-specific stopword lists. We have found that NMF regularly produces more coherent topic descriptors than those generated by the standard approach used with LDA, with higher levels of topic generality and redundancy observed with the latter. It appears that a key role is played by the associated term weighting strategy, where modifications to document term pre-processing and descriptor term post-processing can produce markedly different results.

This evaluation has provided insight into the characteristics and differences between the topic models produced by NMF and LDA. While LDA may offer good general descriptions of broader topics, our results indicate that the higher coherence and lower generality associated with NMF topics mean that the latter method is more suitable when analyzing niche or non-mainstream content. Similarly, although improvements in topic coherence have been found with the use of n-gram terms (Lau, Baldwin, and Newman, 2013), here we have restricted the evaluation to use the common unigram-based approach. Regardless of the topic modeling technique employed, it is clear that close reading of any generated topics is essential.

As certain issues have been raised in relation to coherence measures that are based on individual term pair co-occurrence within a reference corpus (Rosner, Hinneburg, Röder, Nettling, and Both, 2013) (albeit, where NMF not was considered), in future work, we would like to investigate alternative measures. We would also hope to perform a user survey in order to correlate human judgements with our automated results, although this would likely be different to prior coherence studies that requested ratings of individual topics, (Newman, Lau, Grieser, and Baldwin, 2010; Mimno, Wallach, Talley, Leenders, and McCallum, 2011; Lau, Newman, and Baldwin, 2014), where descriptor comparisons would instead need to be considered.

7. Acknowledgements

This research was supported by 2CENTRE, the EU funded Cybercrime Centres of Excellence Network, Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, and the EU FP7 funded VOX-Pol Network of Excellence.

References

Aletras, N. and Stevenson, M. (2013). Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the Tenth International Workshop on Computational Semantics*, IWCS-10, pages 13–22.

Arora, S., Ge, R., Halpern, Y., Mimno, D. M., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2012a). A Practical Algorithm for Topic Modeling with Provable Guarantees. *ArXiv*, abs/1212.4777.

Arora, S., Ge, R., and Moitra, A. (2012b). Learning Topic Models - Going beyond SVD. In *FOCS*, pages 1–10. IEEE Computer Society.

Blei, D. and Lafferty, J. (2009). Topic Models. In *Text Mining: Theory and Applications*. Taylor and Francis.

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems*, NIPS. MIT Press.

Blei, D. M. and Lafferty, J. D. (2006). Correlated Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press.

Blei, D. M., Ng, A. Y., and 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*, GCSL '09.

Boutsidis, C. and Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recogn.*, 41(4):1350–1362.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, NIPS.

Chemudugunta, C., Smyth, P., and Steyvers, M. (2006). Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Advances in Neural Information Processing Systems*, pages 241–248.

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

Gaussier, E. and Goutte, C. (2005). Relation Between PLSA and NMF and Implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 601–602, New York, NY, USA. ACM.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Hoffman, M. D., Blei, D. M., and Bach, F. R. (2010). Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, pages 856–864.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450. ACM.

Lau, J. H., Baldwin, T., and Newman, D. (2013). On Collocations and Topic Models. *ACM Trans. Speech Lang. Process.*, 10(3):10:1–10:14.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, EACL-14, pages 530–539.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for Nonnegative Matrix Factorization. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press.

Lin, C.-J. (2007). Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Comput.*, 19(10):2756–2779.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit, http://mallet.cs.umass.edu.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *ArXiv*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, NIPS.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, PA, USA. Association for Computational Linguistics.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., and Cunningham, P. (2013). An Analysis of Interactions Within and Between Extreme Right Communities in Social Media. In Atzmueller, M., Chin, A., Helic, D., and Hotho, A., editors, *Ubiquitous Social Media Analysis*, volume 8329 of Lecture Notes in Computer Science, pages 88–107. Springer Berlin Heidelberg.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. ACM.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The Author-topic Model for Authors and Documents. In *Proc. 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States. AUAI Press.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., and Both, A. (2013). Evaluating topic coherence measures. In *Advances in Neural Information Processing Systems*, NIPS.

Saha, A. and Sindhwani, V. (2012). Learning Evolving and Emerging Topics in Social Media: A Dynamic NMF approach with Temporal Regularization. In *Proceedings of the 5th ACM international conference on Web search and data mining*, pages 693–702. ACM.

Salton, G. and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Sandhaus, E. (2008). The New York Times Annotated Corpus. Philadelphia: Linguistic Data Consortium.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 952–961, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steyvers, M. and Griffiths, T. (2006). Probabilistic Topic Models. In Landauer, T., Mcnamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Wallach, H. M., Mimno, D., and McCallum, A. (2009a). Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, NIPS.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA. ACM.