Evaluating the Impact of OCR Errors on Topic Modeling

Stephen Mutuvi, Antoine Doucet, Moses Odeo, Adam Jatowt

Evaluating the Impact of OCR Errors on Topic Modeling

Stephen Mutuvi[1], Antoine Doucet[2(✉)], Moses Odeo[1], and Adam Jatowt[3]

[1] Multimedia University Kenya, Nairobi, Kenya
smutuvi@mmu.ac.ke
[2] La Rochelle University, La Rochelle, France
antoine.doucet@univ-lr.fr
[3] Kyoto University, Kyoto, Japan

Abstract. Historical documents pose a challenge for character recognition due to various reasons such as font disparities across different materials, lack of orthographic standards where same words are spelled differently, material quality and unavailability of lexicons of known historical spelling variants. As a result, optical character recognition (OCR) of those documents often yield unsatisfactory OCR accuracy and render digital material only partially discoverable and the data they hold difficult to process. In this paper, we explore the impact of OCR errors on the identification of topics from a corpus comprising text from historical OCRed documents. Based on experiments performed on OCR text corpora, we observe that OCR noise negatively impacts the stability and coherence of topics generated by topic modeling algorithms and we quantify the strength of this impact.

Keywords: Topic modeling · Topic coherence · Text mining
Topic stability

1 Introduction

Recently, there has been rapid increase in digitization of historical documents such as books and newspapers. The digitization aims at preserving the documents in a digital form that can enhance access, allow full text search and support efficient sophisticated processing using natural language processing (NLP) techniques. An important step in the digitization process is the application of optical character recognition (OCR) techniques, which involve translating the documents into machine processable text.

OCR produces its best results from well-printed, modern documents. However, historical documents still pose a challenge for character recognition and therefore OCR of such documents still does not yield satisfying results. Some of

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

M. Dobreva et al. (Eds.): ICADL 2018, LNCS 11279, pp. 3–14, 2018.
https://doi.org/10.1007/978-3-030-04257-8_1

the reasons why historical documents still pose a challenge include font variation across different materials, same words spelled differently, material quality where some documents can have deformations and unavailability of a lexicon of known historical spelling variants [1]. These factors reduce the accuracy of recognition which affects the processing of the documents and, overall, the use of digital libraries.

Among the NLP tasks that can be performed on digitized data is the extraction of topics, a process known as topic modeling. Topic modeling has become a common topic analysis tool for text exploration. The approach attempts to obtain thematic patterns from large unstructured collections of text by grouping documents into coherent topics. Among the common topic modeling techniques are the Latent Dirichlet Allocation (LDA) [3] and the Non-negative Matrix factorization (NMF) [11]. The basic idea of LDA is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words [30]. The standard implementation of both the LDA and NMF rely on stochastic elements in their initialization phase which can potentially lead to instability of the topics generated and the terms that describe those topics [14]. This phenomena where different runs of the same algorithm on the same data produce different outcomes manifests itself in two aspects. First, when examining the top terms representing each topic (i.e. topic descriptors) over multiple runs, certain terms may appear or disappear completely between runs. Secondly, instability can be observed when examining the degree to which documents have been associated with topics across different runs of the same algorithm on the same corpus. In both cases, such inconsistencies can potentially affect the performance of topic models. Measuring the stability and coherence of topics generated over the different runs is critical to ascertain the model's performance, as any individual run cannot decisively determine the underlying topics in a given text corpus [14].

This study examines the effect of noise on unsupervised topic modeling algorithms, through comparison of performance of both the LDA and NMF topic models in the presence of OCR errors. Using a dataset comprising corpus of OCRed documents described in Sect. 4, both the topic stability and coherence scores are obtained and comparison of models' performance on noisy and the corrected OCR text is conducted. To the best of our knowledge, no other study has attempted to evaluate both the stability and coherence of the two models on noisy OCR text corpora.

The remainder of the paper is structured as follows. In Sect. 2 we discuss related work on topic modeling on OCR data. In Sect. 3 we describe the metrics for evaluating the performance of topic models, namely topic stability and coherence, before evaluating LDA and NMF topic models in the presence of noisy OCR in Sect. 4. We discuss the experiment results and conclusion of the paper with ideas for future work in Sects. 5 and 6, respectively.

2 Related Work

2.1 OCR Errors and Topic Modeling

Optical Character Recognition (OCR) enables translation of scanned graphical text into editable computer text. This can substantially improve the usability of the digitized documents allowing for efficient searching and other NLP applications. OCR produces its best results from well-printed, modern documents. Historical documents, however pose a challenge for character recognition and their character recognition does not yield satisfying results. Common OCR errors include punctuation errors, case sensitivity, character format, word meaning and segmentation error where spacings in different line, word or character lead to mis-recognitions of white-spaces [22]. OCR errors may also stem from other sources such as font variation across different materials, historical spelling variations, material quality or language specific to different media texts [1].

While OCR errors remain part of a wider problem of dealing with "noise" in text mining [23], their impact varies depending on the task performed [24]. NLP tasks such as machine translation, sentence boundary detection, tokenization, and part-of-speech tagging on text among others can all be compromised by OCR errors [25]. Studies have evaluated effect of OCR errors on supervised document classification [28,29], information retrieval [26,27], and a more general set of natural language processing tasks [25]. The effect of OCR errors on document clustering and topic modeling has also been studied [9]. The results indicated that the errors had little impact on performance for the clustering task, but had a greater impact on performance for the topic modeling task. Another study explored supervised topic models in the presence of OCR errors and revealed that OCR errors had insignificant impact [31].

While results suggest that OCR errors have small impact on performance of supervised NLP tasks, the errors should be considered thoroughly for the case of unsupervised topic modeling as the models are known to degrade in the presence of OCR errors [9,31]. We thus focus in this work on OCR impacts for unsupervised topics models and in particular on their coherence and stability, and our studies are conducted on large document collection.

2.2 Topic Modeling Algorithms

Topic models aim to discover the underlying semantic structure within large corpus of documents. Several methods such as probabilistic topic models and techniques based on matrix factorization have been proposed in the literature. Much of the prior 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 [2]. One of the most commonly used probabilistic algorithms for topic modeling is the Latent Dirichlet Allocation (LDA) [3]. This is due to its simplicity and capability to uncover hidden thematic patterns in text with little human supervision. LDA represents topics by word probabilities, where words with highest probabilities in each topic

determine the topic. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior. The generative process of LDA is illustrated as follows [3]:

(i) Choose a multinomial topic distribution θ for the document (according to a Dirichlet distribution Dir(α) over a fixed set of k topics)?
(ii) Choose a multinomial term distribution φ for the topic (according to a Dirichlet distribution Dir(β) over a fixed set of N terms)
(iii) For each word position
  (a) Choose a topic Z according to multinomial topic distribution θ.
  (b) Choose a word W according to multinomial term distribution φ.

Various studies have applied probabilistic latent semantic analysis (pLSA) model [4] and LDA model [5] on newspaper corpora to discover topics and trends over time. Similarly, LDA has been used to find research topic trends on a dataset comprising abstracts of scientific papers [2]. Both pLSA and LDA models are probabilistic models that look at each document as a mixture of topics [6]. The models decompose the document collection into groups of words representing the main topics. Several topic models were compared, including LDA, correlated topic model (CTM), and probabilistic latent semantic indexing (pLSI), and it has been found that LDA generally worked comparably well or better than the other two at predicting topics that match topics picked by human annotators [7]. MAchine Learning for LanguagE Toolkit (MALLET) [8] was used to test the effects of noisy optical character recognition (OCR) data using LDA [9]. The toolkit has also been used to mine topics from the Civil War era newspaper dispatch [5], and in another study to examine general topics and to identify emotional moments from Martha Ballards diary [10].

Non-negative Matrix Factorization (NMF) [11] has also been effective in discovering topics in text corpora [12,13]. NMF factors high-dimensional vectors into a low-dimensionality representation. The goal of NMF is to approximate a document-term matrix $\mathbf{A}$ as the product of two non-negative factors $\mathbf{W}$ and $\mathbf{H}$, each with k dimensions that can be interpreted as k topics. Like LDA, the number of topics k to generate is chosen beforehand. The values of $\mathbf{H}$ and $\mathbf{W}$ provide term weights which can be used to generate topic descriptions and topic memberships for documents respectively. The rows of the factor $\mathbf{H}$ can be interpreted as k topics, defined by non-negative weights for each [14].

3 Topic Model Stability

The output of topic modeling procedures is often presented in the form of lists of top-ranked terms suitable for human interpretation. A general way to represent the output of a topic modeling algorithm is in the form of a ranking set containing k ranked lists, denoted $S = R_1, ..., R_k$. The ith topic produced by the algorithm is represented by the list R_i, containing the top t terms which are most characteristic of that topic according to some criterion [15].

The stability of a clustering algorithm refers to its ability to consistently produce similar results on data originating from the same source [16]. Standard implementations of topic modeling approaches, such as LDA and NMF, commonly employ stochastic initialization prior to optimization. As a result, the algorithms can achieve different results on the same data or data drawn from the same source, between different runs [14]. The variation manifests itself either in relation to term-topic associations or document-topic associations. In the former, the ranking of the top terms that describe a topic can change significantly between runs. In the latter, documents may be strongly associated with a given topic in one run, but may be more closely associated with an alternative topic in another run [14].

To quantify the level of stability or instability present in a collection of topic models $\{M_1, ..., M_r\}$ generated over r runs on the same corpus, the Average Term Stability (ATS) and Pairwise Normalized Mutual Information (PNMI) measures have been proposed [14].

We begin by determining the Term Stability (TS) score, which involves comparison of the similarity between two topic models based on a pairwise matching process. The measuring of the similarity between a pair of individual topics represented by their top t terms is based on the Jaccard Index:

$$Jac(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} \tag{1}$$

where R_i denotes the top t ranked terms for the i-th topic (topic descriptor). We can use Eq. 1 to build a measure of the agreement between two complete topic models (i.e., Term Stability):

$$TS(M_i, M_j) = \frac{1}{k} \sum_{x=1}^{k} Jac(R_{ix}, \pi(R_{ix})) \tag{2}$$

where $\pi(R_{ix})$ denotes the topic in model M_j matched to R_{ix} in model M_i by the permutation π. Values for TS take the range $[0, 1]$, where similarity between two topic models will result in a score of 1 if identical.

For a collection of topic models M_r, we can calculate the Average Term Stability (ATS):

$$ATS = \frac{1}{r \times (r-1)} \sum_{i,j:i \neq j}^{r} TS(M_i, M_j) \tag{3}$$

where a score of 1 indicates that all pairs of topic descriptors matched together across the r runs contain the same top t terms [14].

Topic model stability can also be established from document-topic associations. PNMI determines the extent to which the dominant topic for each document varies between multiple runs. The overall level of agreement between a set of partitions generated by r runs of an algorithm on the same corpus can be computed as the mean Pairwise Normalized Mutual Information (PNMI) for all

pairs:

$$PNMI = \frac{1}{r \times (r-1)} \sum_{i,j:i \neq j}^{r} NMI(P_i, P_j) \tag{4}$$

where P_i is the partition produced from the document-topic associations in model M_i. If the partitions across all models are identical, PNMI will yield a value of 1.

3.1 Quality of Topics

While topic model stability is important, it is unlikely to be useful without meaningful and coherent topics [14]. Measuring topic coherence is critical in assessing the performance of topic modeling approaches in extracting comprehensible and coherent topics from corpora [17]. The intuition behind measuring coherence is that more coherent topics will have their top terms co-occurring more often together across the corpus. A number of approaches for evaluating coherence exist, although many of these are specific to LDA. A more general approach is the Topic Coherence via Word2Vec (TC-W2V), which evaluates the relatedness of a set of top terms describing a topic [18]. TC-W2V uses the increasingly popular word2vec tool [21] to compute a set of vector representations for all of the terms in a large corpus. The extent to which the two corresponding terms share a common meaning or context (e.g. are related to the same topic) is assessed by measuring the similarity between pairs of term vectors. Topics with descriptors consisting of highly-similar terms, as defined by the similarity between their vectors, should be more semantically coherent [19].

TC-W2V operates as follows. The coherence of a single topic t_h represented by its t top ranked terms is given by the mean pairwise cosine similarity between the t corresponding term vectors generated by the *word2vec* model [18].

$$\text{coh}(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} cosine(wv_i, wv_j) \tag{5}$$

An overall score for the coherence of a topic model T consisting of k topics is given by the mean of the individual topic coherence scores:

$$\text{coh}(T) = \frac{1}{k} \sum_{h=1}^{k} \text{coh}(t_h) \tag{6}$$

In the next section, we use the theory described in this section to determine the stability and coherence scores of topics generated by LDA and NMF topic models, from the data described in Sect. 4.1.

4 Experiments

In this section, we seek to apply topic modeling techniques, LDA and NMF, on the OCR text corpus described below, in an attempt to answer the following two questions:

(i) To what extent do OCR errors affect the stability of topic models?
(ii) How do the topic models compare in terms of topic coherence, in the presence of OCR errors?

4.1 Data Source

A large corpus of historical documents [20], comprising twelve million OCRed characters along with the corresponding Gold Standard (GS) was used to model topics. This dataset comprising monographs and periodical has an equal share of English- and French-written documents ranging over four centuries. The documents are sourced from different digital collections available, among others, at the National Library of France (BnF) and the British Library (BL). The corresponding GS comes both from BnF's internal projects and external initiatives such as Europeana Newspapers, IMPACT, Project Gutenberg, Perseus, Wikisource and Bank of Wisdom [20].

4.2 Experimental Setup

The experiment process involved applying LDA and NMF topic models to the noisy OCRed_toInput, aligned OCRed and Gold Standard (GS) text data. Only the english language documents from the dataset was considered in the experiment. The OCRed_toInput is the raw OCRed text while the aligned OCR and GS represent the corrected version of the text corpus provided for training and testing purposes. The alignment was made at the character level using "@" symbols with "#" symbols corresponding to the absence of GS either related to alignment uncertainties or related to unreadable characters in the source document [20].

The three categories of text were extracted from the corpus, separately preprocessed and the models were applied on each one of them to obtain topics. Fifty topic models (M_{50}) for each value of k, where k is the number of topics ranging from 2 to 8, were generated for both the NMF and LDA. The selection of this number of topic k was based on a previous study which proposed an approach for choosing this parameter using a term-centric stability analysis strategy [15]. The LDA algorithm was implemented using the popular Mallet implementation with Gibbs sampling [8].

The stability measures for the two topic modeling techniques were obtained and evaluated to determine their performance on the noisy and corrected OCR text. A high level of agreement between the term rankings generated over multiple runs of the same model is an indication of high topic stability [15]. The assumption in this study was that noisy OCRed text would register a lower topic stability value compared to the corrected text, an indication that OCR errors have a negative impact on topic models.

4.3 Evaluation of Stability

To assess the stability of topics generated by each model, the term-based measure (ATS) for each topic's top 10 terms and the document-level (PNMI) measure

were computed using Eqs. 3 and 4, respectively. The results of the average topic stability and the average partition stability are shown in Tables 1 and 2, respectively. Figure 1 provides a graphical representation of LDA and NMF stability scores on the different dataset categories.

Table 1. Average topic stability.

Model	Dataset	Mean stability
LDA	GS_aligned	0.265*
LDA	OCR_aligned	0.256
LDA	OCR_toInput	0.252*
NMF	GS_aligned	0.414*
NMF	OCR_aligned	0.384
NMF	OCR_toInput	0.383*

Asterisk (*) indicates p-value was less than 0.05 for independent samples t-test

Both models recorded higher average topic stability on the aligned text compared to the raw OCR text. The mean stability on the Gold Standard text was 0.265 and 0.414 while for the noisy OCR text was 0.252 and 0.383 for LDA and NMF topic models, respectively.

Table 2. Average partition stability.

Model	Dataset	Mean stability
LDA	GS_aligned	0.115
LDA	OCR_aligned	0.115
LDA	OCR_toInput	0.114
NMF	GS_aligned	0.117*
NMF	OCR_aligned	0.115
NMF	OCR_toInput	0.114*

Asterisk (*) indicates p-value was less than 0.05 for independent samples t-test

The average partition stability for LDA remained unchanged for both aligned and raw OCR text. However, NMF recorded a mean partition stability score of 0.117 and 0.114 for the Gold Standard and raw OCR text, respectively.

4.4 Topic Coherence Evaluation

The quality of the topics generated by the models was evaluated by computing the coherence of the topic descriptors using the approach described in Sect. 3.1.

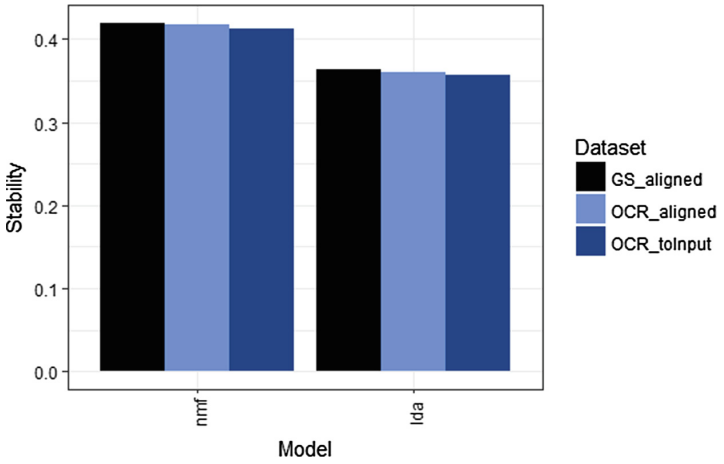

Fig. 1. Model stability on noisy and corrected OCR texts.

The results of the average coherence score for LDA and NMF algorithms, on the noisy and corrected data are presented in Table 3.

Table 3. Mean topic coherence.

Model	Dataset	Mean coherence
LDA	GS_aligned	0.3622
LDA	OCR_aligned	0.3585
LDA	OCR_toInput	0.3529
NMF	GS_aligned	0.4748
NMF	OCR_aligned	0.4737
NMF	OCR_toInput	0.4720

The mean coherence score on the aligned OCR text was 0.4737 and 0.3585 for NMF and LDA algorithms respectively. On the other hand, the mean coherence using the raw OCR text was marginally lower recording 0.4720 for NMF and 0.3529 for LDA topic model.

5 Discussions

Topic modeling algorithms have been evaluated based on model quality and stability criteria. The quality and stability of the algorithms was determined by examining topic coherence and term and document stability respectively. Overall, the aligned corpus text had higher stability score compared to the noisy

OCR input text for both the LDA and NMF topic modeling techniques. The NMF algorithm yielded the most stable results both at the term and document-level, as shown in Fig. 1.

On the other hand, the evaluation of topic coherence showed topics from the corrected corpus were more coherent compared to the original noisy text for both the LDA and NMF models. As expected, LDA had a lower coherence score than NMF, which may reflect the tendency of LDA to produce more generic and less semantically-coherent terms [18]. The difference in average coherence score between the models was relatively small for the aligned OCR and noisy OCR text corpus.

6 Conclusions

It is evident from this study that OCR errors can have a negative impact on topic modeling, therefore affecting the quality of the topics discovered from text datasets. Overall, this can impede the exploration and exploitation of valuable historical documents which require use of OCR techniques to enable their digitization. Advanced OCR post correction techniques are required to address the impact of OCR errors on topic models.

Future research can explore the impact of OCR errors on the accuracy of other text mining tasks such sentiment analysis, document summarization and named entity extraction. In addition, multi-modal text mining approaches that put into consideration textual and visual elements can be explored to determine their suitability in processing and mining of historical texts. Evaluating the stability and coherence for different number of topic models can also be examined further.

References

1. Silfverberg, M., Rueter, J.: Can morphological analyzers improve the quality of optical character recognition? In: Septentrio Conference Series, vol. 2, pp. 45–56 (2015)
2. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487–494. AUAI Press (2004)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Newman, D.J., Block, S.: Probabilistic topic decomposition of an eighteenth-century American newspaper. J. Assoc. Inf. Sci. Technol. **57**(6), 753–767 (2006)
5. Nelson, R.K.: Mining the dispatch (2010)
6. Yang, T.I., Torget, A.J., Mihalcea, R.: Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities, pp. 96–104. Association for Computational Linguistics (2011)
7. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)

8. McCallum, A.K.: Mallet: a machine learning for language toolkit (2002)
9. Walker, D.D., Lund, W.B., Ringger, E.K.: Evaluating models of latent document semantics in the presence of OCR errors. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 240–250. Association for Computational Linguistics (2010)
10. Blevins, C.: Topic modeling Martha Ballard's diary. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary. Accessed 23 Feb 2018
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–91 (1999)
12. Arora, S., Ge, R., Moitra, A.: Learning topic models - going beyond SVD. In: Proceedings of 53rd Symposium on Foundations of Computer Science, pp. 1–10. IEEE (2012)
13. Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Celebi, M.E. (ed.) Partitional Clustering Algorithms, pp. 215–243. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09259-1_7
14. Belford, M., Mac Namee, B., Greene, D.: Stability of topic modeling via matrix factorization. Expert Syst. Appl. **91**, 159–169 (2018)
15. Greene, D., O'Callaghan, D., Cunningham, P.: How many topics? Stability analysis for topic models. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS (LNAI), vol. 8724, pp. 498–513. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44848-9_32
16. Lange, T., Roth, V., Braun, M.L., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural Comput. **16**(6), 1299–1323 (2004)
17. Fang, A., Macdonald, C., Ounis, I., Habel, P.: Using word embedding to evaluate the coherence of topics from Twitter data. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2016, pp. 1057–1060 (2016)
18. O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Syst. Appl. **42**(13), 5645–5657 (2015)
19. Greene, D., Cross, J.P.: Exploring the political agenda of the European parliament using a dynamic topic modeling approach. Polit. Anal. **25**, 77–94 (2017)
20. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.P.: Impact of OCR errors on the use of digital libraries: towards a better access to information. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (2017)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
22. Afli, H., Barrault, L., Schwenk, H.: OCR error correction using statistical machine translation. In: 16th International Conference Intelligent Text Processing Computational Linguistics (CICLing 2015), vol. 7, pp. 175–191 (2015)
23. Knoblock, C., Lopresti, D., Roy, S., Subramaniam, V.: Special issue on noisy text analytics. Int. J. Doc. Anal. Recogn. **10**(3–4), 127–128 (2007)
24. Eder, M.: Mind your corpus: systematic errors in authorship attribution. Literary Linguist. Comput. **10**, 1093 (2013)
25. Lopresti, D.: Optical character recognition errors and their effects on natural language processing. Presented at The Second Workshop on Analytics for Noisy Unstructured Text Data, Sponsored by ACM (2008)
26. Taghva, K., Borsack, J., Condit, A.: Results of applying probabilistic IR to OCR text. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 202–211. Springer, New York (1994)

27. Beitzel, S., Jensen, E.C., Grossman, D.A.: A survey of retrieval strategies for OCR text collections. In: Proceedings of 2003 Symposium on Document Image Understanding Technology (2003)
28. Taghva, K., Nartker, T., Borsack, J., Lumos, S., Condit, A., Young, R.: Evaluating text categorization in the presence of OCR errors. In: Document Recognition and Retrieval VIII. International Society for Optics and Photonics, vol. 4307, pp. 68–75 (2000)
29. Agarwal, S., Godbole, S., Punjani, D., Roy, S.: How much noise is too much: a study in automatic text classification. In: Proceedings of the Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 3–12 (2007)
30. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Handbook of Latent Semantic Analysis, vol. 427, no. 7, pp. 424–440 (2007)
31. Walker, D., Ringger, E., Seppi, K.: Evaluating supervised topic models in the presence of OCR errors. In: Document Recognition and Retrieval XX, vol. 8658, p. 865812. International Society for Optics and Photonics (2013)