Finding topic boundaries in literary text

Heyns, Nuette
North-West University, South Africa
nuette.heyns@gmail.com

van Zaanen, Menno
South African Centre for Digital Language Resources, South Africa
menno.vanzaanen@nwu.ac.za

Abstract

When performing a distant reading analysis of large amounts of literary texts, we would like to be able to automatically identify the high level structure or story lines of these texts. Story lines are not always linear, but contain transitions, such as flashbacks or changes of scenery. While working towards our goal of identifying story lines in text, we first start by identifying topic transitions. We propose a system that aims to identify a boundary describing a topic transition in the text. First, we split the text in short snippets. Next, topics are assigned to each of the snippets using LDA, a topic modelling approach. Based on this sequence of LDA topics, potential transition boundaries between snippets are identified. Potential transitions occur between snippets with the smallest intersection of the LDA topics that occur on either side of the potential transition. If multiple potential transitions are available, the system selects one at random. To evaluate this system, we apply it to the concatenation of two texts such that the real boundary is known. We provide results of this system with respect to a random baseline and an oracle system that always selects the best transition when more than one possible transition is available. The system consistently outperforms the baseline. Future work will focus on extending this system to allow for the identification of multiple transitions.

Keywords: topic modelling, LDA, boundary identification

1 Introduction

With the availability of huge amounts of texts, in depth literary analysis of all texts using manual close reading approaches is infeasible. Distant reading approaches (Moretti 2013) that rely on the automatic analysis of the texts should be considered instead. The idea of distant reading is that the computer can perform large scale and objective analyses of the texts, in contrast to the more time consuming and subjective manual analyses. (However, it is generally assumed that close reading approaches can provide a more fine-grained analysis compared to the distant reading approaches.)

One type of literary analysis deals with the identification of story lines, that can be found, for instance, in literary texts. Structuralist theorist Gérard Genette discerns four important levels of a literary text; order, duration, frequency and mood (Genette et al. 1980). We focus on the first level, order, where the sequence of events is viewed in relation to the order of narration. Many literary texts do not follow a linear story line, but apply literary techniques such as the use of different perspectives, different locations, or variations in the time line (e.g., flashbacks or flashforwards). In particular, we are interested in the transitions that occur in the story lines throughout a literary text. This allows for high level comparisons, for instance, of writing styles of different authors or structural differences in texts from different genres.

Transitions in the story line can be seen as boundaries, separating the text into parts of the text that have different properties. How these parts are different depends on the type of transition, but because the text before the transition and that after will be different in some aspect(s), we may assume that such transitions can be automatically identified based on the differences between properties of the part of the text before and after the transition.

In this article, we propose a method that aims to automatically identify a topic transition in a text. This method assumes that transitions can be identified by considering changes that can be described by

topics (as identified using the Latent Dirichlet Allocation, or LDA (Blei et al. 2003*a*), topic model). In particular, we subdivide the text into smaller snippets and apply LDA to these snippets to determine their topics. The method then analyses the sequence of LDA topics to identify potential transitions between snippets. These potential transitions occur at all positions where the intersection of the sets of LDA topics that occur before and the LDA topics that occur after the potential transition is the smallest. In other words, the system finds boundaries such that the topics occurring "on the left" of the boundary is maximally different from the topics that occur "on the right" of the boundary. These are positions where the text before and after the transition is different on the basis of LDA topics.

To evaluate the method, we construct a text by concatenating two different texts, such that the position of the real transition is known. We then apply our method, which proposes the location of a transition. The proposed transition is then compared against the real transition. To measure how well the proposed transition fits the real transition, the root mean squared error (RMSE) is computed, which takes distance into account. Lower values for RMSE are better. This system is evaluated against a baseline (which does not use the LDA model) and an oracle system (which always selects the best possible transition, in contrast to the proposed system which makes a selection from all possible transitions at random).

The LDA topic modelling system has a parameter that indicates how many topics LDA may assign to the snippets. As the system relies on the differences between the topics on both sides of the potential transition, we may expect that the number of LDA topics will have an influence of the performance of the transition identification system. In fact, in order to apply the system to the snippets of the text, we need to define the number of LDA topics beforehand, so it is useful to know more about the influence of the number of LDA topics on the performance of the system to make an informed choice when applying the system to a new text.

In this article we will focus on the following research questions.

1. Can a system that identifies a transition in a text based on LDA topics of snippets outperform a random baseline?
2. What is the influence of the random selection of the possible boundaries on the performance of the LDA based system?
3. What is the influence of the number of LDA topics on the performance of the LDA based system?

To answer the first question, we will apply the system and the random baseline to a text with a known transition and compare the results. For the second question, we compare the results of the system against an oracle system, which always selects the best of the possible boundaries. We also run the system with several values for the number of LDA topics and evaluate their performance to better understand how to answer the third question.

2 Background

The system proposed in this article depends heavily on the performance of LDA. Fortunately, there has already been research on the performance of LDA in different settings. In particular, the length of the documents given to LDA has a direct influence of the performance of LDA. We will look at this research first. Next, we briefly discuss different ways of evaluating the performance of LDA, mostly focusing on the limitations of evaluating LDA directly.

With respect to the automatic identification of transitions in literary text, unfortunately, to our knowledge there is not much previous research. Aurnhammer et al. (2019) performed a comparison between a close reading approach, which relied on manually annotated Reddit posts and a distant reading approach which relied on the identification of topics using LDA. Here the texts were already separated (as they were individual posts), but this work showed that there is a relationship between manually annotated texts and LDA texts. Similarly,

for instance, Huang & Huant (2013) and Zhou et al. (2016) genereated story lines given a collection of news articles. Gupta et al. (2009) aimed to visualise story lines (from video data), but rely on weakly labelled data.

2.1 Length of LDA documents

Sbalchiero & Eder (2020) focused on the model fitting process of topic modelling when applied to long texts. In their study, they examined the performance of LDA on literature text by splitting the text using six different sample sizes (500, 1000, 5000, 10000, 20000, and 50000). Based on this, they found that there is a relationship between the length of text chunks and the number of topics. Intuitively an extremely short text chunk and a large number of topics will provide many very specific topics, which causes the model to overfit. Reversely, an extremely large text chunk combined with too few topics will result in very broad, general topics causing the model to underfit. Sbalchiero & Eder (2020) state that given a corpus, the optimal number of topics is inversely proportional to the length of text chunks. Therefore, the larger the size of the text chunk, the lower the optimal number of topics will be. However, they also mention that the extreme cases where there are too many topics combined with a very short sample chunk will overfit the model and too few topics combined with an extremely large text chunk will underfit the model. From this statement, we can derive that the optimal number of topics to size of the text chunk should be in equilibrium. Sbalchiero & Eder (2020) conclude that the best number of topics for different sizes of the samples should be evaluated using, for example, the elbow method suggested by Kodinariya & Makwana (2013).

According to Sbalchiero & Eder (2020), previous studies have already demonstrated that LDA performs well when applied to short texts, but there is a lack of empirical evidence to show that LDA also performs well on longer texts. Syed & Spruit (2017) argue that longer text are less affected by noise in the topic-word distributions, resulting in more coherent topics. However, limited research has been done on this subject.

Jockers & Mimno (2013) indicate that the ideal size of the sample texts should be large enough to allow for the proper measurement of word cooccurences, but small enough that it can reasonably be assumed to contain a small number of themes. They found that applying LDA to full texts typically results in vague topics. However, splitting texts into approximately 1000 word samples, breaking at the nearest sentence boundary, results in more highly interpretable topics. Studies like Syed & Spruit (2017), Blei et al. (2003*b*) and others suggest using abstracts as a suitable size of sample texts.

2.2 Evaluation of LDA

LDA models can be evaluated using either extrinsic or intrinsic methods. Extrinsic evaluation methods measure LDA models' performance on a secondary task, such as document classification or information retrieval (Wallach et al. 2009). Intrinsic methods include measurements that help distinguish between topics that are semantically interpretable and topics that are artefacts of statistical inference. Usually an intrinsic method rely on the estimation of the probability of an unseen held-out data set given the trained model (Wallach et al. 2009). Popular intrinsic methods are log-likelihood and perplexity measures, as well as topic coherence.

The log-likelihood approach measures how well an LDA model fits the data. The probability of a held-out data set, not used during training, can be estimated in several ways, such as importance sampling methods, harmonic mean, annealed importance sampling, a Chib-style estimator, or a left-to-right evaluation algorithm (Wallach et al. 2009). Perplexity can also be used to measure the quality of the LDA model. Perplexity describes how well an LDA model predicts a topic for a sample by computing the normalised log-likelihood of a held-out test set. A model will be considered good when it has a high log-likelihood and, hence, a low perplexity score. Chang et al. (2009) have, however, shown that the log-likelihood and perplexity scores

have poor correlation to human judgement and are sometimes even slightly anti-correlated.

Another approach deals with the top words found in a topic. For each word, a vector representation can be created based on occurrences in large amounts of texts. Based on these vectors, the LDA topics can be evaluated by measuring the cosine distance between the words that describe the topic. If words from the same topic are closer together, the coherence is considered high. The underlining idea behind topic coherence is the distributional hypothesis of linguistics. The distributional hypothesis states that words with similar meaning tend to occur in similar contexts (Harris 1954).

Note that, ideally, human topic rankings should be available, to compare the LDA topic coherence scores to the human topic rankings. Unfortunately, in most cases topics identified by humans are not available and researchers have to rely on some automatically computed coherence score alone.

3 Methodology

3.1 Systems

The transition identification system that we propose in this article consists of four steps. First, it takes the input text T and subdivides the text into n snippets: $S = \langle s_1, \ldots, s_n \rangle$, where $T = s_1 \oplus s_2 \oplus \cdots \oplus s_n$ with $\oplus$ the concatenation operator. Second, this sequence of snippets (S) is given to the LDA system, which essentially provides a mapping LDA, which results in a sequence of LDA topics: $LDA(S) = \langle LDA(s_1), LDA(s_2), \ldots, LDA(s_n) \rangle$. Third, potential transitions are identified. Each position between two snippets, (s_x, s_{x+1}) in the sequence (with $x = 1 \ldots n-1$) is considered. For each of these positions, the size of the intersection of the set of LDA topics before this position and the set of LDA topics after the position is computed. The minimum value of all of these intersections indicates the best potential transition and there may be several positions that have the same minimum intersection sizes: $\arg\min_{x=1}^{n} | \bigcup_{i=1}^{x} LDA(s_i) \cap \bigcup_{j=x+1}^{n} LDA(s_j) |$. Finally, the system selects one of the potential transitions. If there are multiple potential transitions, it selects one at random.

The transition identification system is compared to two other systems: a baseline and an oracle system. The baseline system does not use any LDA information, but selects a transition at random from all possible positions between the snippets. This system serves as a lower limit. In contrast, the oracle system follows the regular transition identification system with one difference: when multiple potential transitions are identified, this system selects the best of these potential transitions. In other words, it makes use of information of where the real transition can be found. This method serves as an upper limit.

3.2 Data

In order to properly evaluate the performance of the system, we need to apply the system to a text in which the transition is known. For this, we create a text by concatenating two source texts that we know discuss different topics. Here, we used two books as source texts: Utilitarianism (Mill 1861) and Hide and Seek (Collins 1861). Straightforward preprocessing is applied to these text: stopwords are removed using NLTK[1], as these words occur so frequently that they do not help in identifying LDA topics of the snippets (but they do have an impact on the size of the snippets). Additionally, the text is lower cased, lemmatised, and punctuation is removed using spaCy[2].

From these two books, we selected the first 25 snippets of 500 words each, resulting in a list of 50 snippets in total with the known transition after 25 snippets. Table 1 shows a sample from both of the source texts.

3.3 Experimental settings

As mentioned before, the transition identification system relies on LDA to identify topics for each of the snippets. LDA has a parameter that sets the number of topics that LDA is allowed to assign to the snippets. As this is a manually assigned param-

Table 1: Sample from each of the two source texts.

Source text	Sentence:
Mill (1861)	desire different thing desire happiness love music desire health They included happiness They elements desire happiness made Happiness abstract idea concrete whole parts And utilitarian standard sanctions approves Life would poor thing ill provided sources happiness provision nature things originally indifferent conducive otherwise associated satisfaction primitive desires…
Collins (1861)	ruddy face suddenly turned pale left circus determined find really going behind red curtain He walked round outside building wasting time found door apply admission At last came sort passage tattered horse-cloths hanging outer entrance You can't come said shabby lad suddenly appearing inside shirt sleeves Mr. Blyth took half-a-crown I want see deaf dumb child directly Oh right go muttered lad pocketing money greedily Valentine hastily entered passage As soon inside sound reached ears heart sickened turned faint No words describe horror helplessness moan pain dumb human creature…

eter, we can vary this parameter in the experiments. In the experiments described in this article, we varied the number of LDA topics from two to 30 in steps of two. For each of the number of LDA topics, the system is run 100 times (as LDA may lead to slightly different results due to a random factor.) We provide the median, average, and standard deviation results for each of these settings.

3.4 Evaluation

To measure how well the different systems perform, we need to decide on an evaluation metric. We are interested in finding a transition that is as close as possible to the real transition (the real transition is known as we have essentially created a text by concatenating two different texts). In other words, we would like to have an evaluation metric that takes into account the distance between the proposed and real transition. For this, we use the root mean squared error, which is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(p_i - r\right)^2}{n}}$$

where n is the number of runs, p_i is the position of the proposed transition position (which can range from one to 49) in run i (which ranges from one to 100, as we run the system 100 times due to the random factor of LDA and the random selection in case of multiple possible transitions) and r is the position of the real transition (at position 25). The scikit-learn Python package[3] was used to calculate the RMSE.

Note that this approach does not directly evaluate the performance of LDA, but instead focuses on how well the overall system identifies the boundaries. In other words, we perform an extrinsic evaluation.

4 Results

To investigate the performance of the transition identification system, we provide the RMSE results of the system as well as the random baseline and oracle system in Table 2. This table also shows this information for each of the settings for the number of LDA topics.

From these results we see that the RMSE of the baseline is around 15. Note that the baseline always selects a random position for the transition, which may range from one to 49, with the real transition at position 25.

Our system performs perfectly with two LDA topics as can be seen by the RMSE of 0.0 and a standard

Table 2: RMSE results of each system for the range of LDA topics. Note that the baseline does not rely on LDA and hence has no # LDA topics provided.

system	# topics	median	mean	sd
Baseline		15.0	14.603	1.067
Our	2	0.0	0.000	0.000
Our	4	0.0	0.957	2.941
Our	6	0.0	2.943	5.357
Our	8	0.0	5.814	7.437
Our	10	6.0	7.229	7.390
Our	12	6.5	9.543	8.576
Our	14	6.0	8.371	9.305
Our	16	9.0	10.057	8.485
Our	18	9.0	10.671	8.759
Our	20	12.0	12.914	8.165
Our	22	9.5	12.114	9.454
Our	24	10.5	12.086	7.910
Our	26	10.5	11.014	8.893
Our	28	10.0	12.886	8.596
Our	30	10.0	12.129	9.119
Oracle	2	0.0	0.000	0.000
Oracle	4	0.0	0.471	2.263
Oracle	6	0.0	0.843	3.242
Oracle	8	0.0	1.843	4.652
Oracle	10	0.0	3.429	6.788
Oracle	12	0.0	3.000	6.347
Oracle	14	0.0	2.514	6.611
Oracle	16	0.0	1.243	4.206
Oracle	18	0.0	1.543	4.989
Oracle	20	0.0	2.429	6.135
Oracle	22	0.0	1.700	5.176
Oracle	24	0.0	0.714	2.649
Oracle	26	0.0	1.714	5.491
Oracle	28	0.0	1.229	4.304
Oracle	30	0.0	0.914	3.202

deviation of 0.0 (remember, lower values of RMSE are better as they relate to the distance of the position of the proposed transition compared to the position of the real transition). In each run, exactly the right position for the transition is identified. Effectively, LDA identifies that there are two main topics that can be identified in the complete text and these correspond to the two original texts that were concatenated.

The performance of our system gradually deteriorates when more LDA topics are made available. When four LDA topics are available, the performance is still quite good with a RMSE of 0.957, but the standard deviation is already 2.941, which indicates that if a wrong transition is identified it may be relatively far away from the real position.

Increasing the number of LDA topics generally decreases the performance. Overall, the mean RMSE becomes larger, indicating that more often incorrect positions for the transition are proposed. The standard deviation also becomes relatively large, which again indicates the spread of proposed transitions. Note that the median also becomes larger which emphasises the larger spread. The slight improvement of the system at 22 LDA topics is probably due to the random factors of LDA and the selection of the proper transition. The standard deviation is relatively large, so it is unlikely to be a real improvement.

If we now consider the performance of the oracle system, we see that the oracle system, like our system, performs well with low number of available LDA topics. The fact that our system already performed perfectly with two LDA topics means that the oracle system cannot improve as our system already always selects the best position for the transition. However, with four available LDA topics, sometimes the oracle system leads to runs that do not contain the correct transition. Here we can see the impact of the random factor of LDA as the oracle system always selects the best transition position. Incorrect possible transitions are also found, which leads to a lower score for our system with four LDA topics. The performance of the oracle system

also deteriorates with larger number of LDA topics, which means that the correct transition cannot be found in any of the proposed transitions according to the sequence of LDA topics. However, the median remains at 0, indicating that often the correct transition is proposed. There are runs in which the random factor of LDA leads to sets of possible transitions where the correct transition is not proposed. In other words, the lower performance of the oracle system with larger number of LDA topics is the result of the random factor in the LDA system, whereas the difference between the oracle system and our system can be attributed to the random selection of transitions when multiple possible transitions are identified.

To support the idea that the number of potential transitions increases with the number of LDA topics, we can look at the correlation between the number of LDA topics and the number of possible boundaries. Computing Pearson's product-moment correlation results in a moderate significant ($p < .0001$) correlation between the number of LDA topics and the number of possible boundaries with $r = .698$ (an $r > .70$ is considered a strong correlation). Figure 1 shows the relationship between the number of topics and the number of possible boundaries the system identifies. The x-axis of the graph shows the number of LDA topics and the y-axis the number of potential transitions identified by the system. Note that the transparency of the points in the graph indicate how frequently that situation occurs, with darker points having higher frequency. We see that when increasing the number of LDA topics indeed increases the number of possible positions for the transitions. This results in situations where our transition identification system has a harder time as there are more possible transitions to choose from. The line is computed using the local polynomial regression fitting and the shaded area indicates the 95% confidence interval.

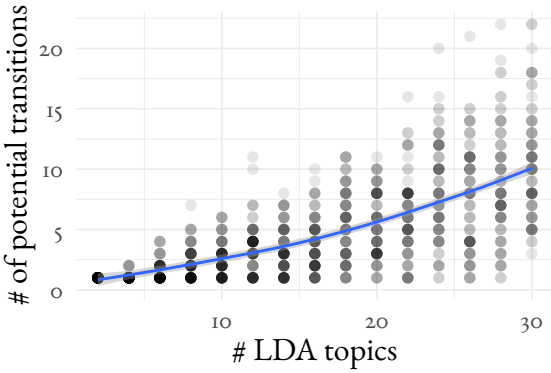

Figure 1: The relationship between the number of LDA topics and the number of possible transitions proposed by the system. Darker points indicate higher frequency of that situation. The line indicates the local polynomial regression fitting and the shaded area around the line represents the 95% confidence interval.

5 Discussion

Ultimately, we are interested in the transitions that occur in the story lines throughout a literary text. However, given the nature of story lines, this task is difficult to achieve. A more contained problem is that of identifying transitions in the topics of literary text. There should be no argument of where the topic transitions occur in the text, therefore the algorithm can be evaluated against a clear answer. We propose to build on this algorithm in the future so that it can also identify story line transitions.

In this article, we proposed a system that aims to identify topic transitions in the story line in a text by first subdividing the text into smaller snippets. This sequence of snippets is used as the input to LDA, which assigns topics to each of the snippets. Next, based on the size of the intersection of the LDA classes of the snippets "to the left" and "to the right" of each of the positions between snippets, the best potential transitions are identified. If multiple potential transitions are found, one is selected at random.

The reasoning behind using the size of the intersection of the LDA topics on both sides of the potential transition is that transitions will show a change of topics. The current system assumes that the topics at one side of the transition will not occur at the other side of the transition (or at least less frequently). This means that the approach described here relies on the performance of LDA in assigning

the correct topics.

We have evaluated the system with snippets from two texts. One may assume that it would be easiest to identify the transition between the texts if only two LDA topics are requested. The results show that this is indeed true. However, it may be the case that the snippets from one source text already contain two or more topics. In that case, LDA may have problems assigning the right topics to the snippets. Essentially, in that case, underfitting will occur. This corresponds to the idea described by Sbalchiero & Eder (2020).

It is interesting, however, that the performance goes down if the number of LDA topics goes up. The system does not directly evaluate the performance of the LDA system, it only relies on the intersection of the topics. From this, we can conclude that increasing the number of LDA topics results in the creation of topics that occur frequently on both sides of the potential transitions, which essentially introduces noise when trying to decide on the best transition. This again, shows that the system is overfitting the data, again following the results from Sbalchiero & Eder (2020).

Based on the results, we see that the proposed system can be used to identify transitions in a text. The system is relatively stable, even if the system tries to assign more LDA topics than are represented in the text, the system still has reasonable performance. Currently, however, several variables have not been evaluated yet, such as the influence of the actual texts, and the length of the snippets. We already know (again, based on Sbalchiero & Eder (2020)) that there is a relationship between these variables.

6 Conclusion

In this article, we aimed to answer three related research questions. The first question focused on the performance of the transition identification system that we introduced in this article. This system subdivides a longer text into smaller snippets, which are the input to LDA. The system then tries to identify possible transitions by considering the size of the intersection of the LDA topics on either side of the possible transition, which may occur between each pair of snippets. The positions that show the smallest intersection are considered possible transitions and if more than one is found, the system selects one at random.

The system consistently outperforms the baseline, indicating that the information that comes from LDA is indeed useful. When more LDA topics are requested, the performance goes down, but perfect results were found when LDA was run with only two topics.

The second question dealt with the influence of the random selection of the system in case multiple transitions were found. We saw that an oracle system, which always selects the best transition, leads to somewhat better results, but even with the oracle system, the performance drops when using more LDA topics. Sometimes the oracle system does lead to perfect results and sometimes it does not, which is the influence of the random factor in the LDA system.

The third question focused on the influence of the number of LDA topics the system used. We see from the result that increasing the number of LDA topics leads to lower results. This means that with higher numbers of LDA topics, additional topics that do not really seem to describe proper topics are assigned to snippets in the text. We can conclude this as they influence the performance of the system as more topics can be found on both sides of the potential transitions. Essentially, this introduces more noise in the LDA topics, due to overfitting.

7 Future work

The research described in this article shows good results, but also raises questions that should be addressed in future work. Specifically, we identify three main areas for future work.

First, the current system only identifies one transition in a text. Future work will need to focus on extending the system to allow for the identification of multiple transitions. The same evaluation strategy can be taken as it is possible to concatenate

three or more texts together. However, the evaluation metric will need to be adjusted to handle multiple boundaries. This scenario, however, is closer to the scenario we would find with a real text. It is yet unclear, exactly how the identification of the transitions will then need to take place. Perhaps a probabilistic approach which assigns probabilities for each of the positions between snippets, combined with a threshold may work. It is also unclear what the influence of the number of LDA topics will be.

Second, the current experiments were only performed on snippets from one pair of texts. Some of the specific results we found (such as the drop in performance around 22 topics) may be attributed to those texts. Experiments on additional pairs of texts, for instance, closer related semantically, may provide more insight in the actual behaviour of the system.

Finally, We may want to investigate the influence of the length of the snippets that are being used when assigning the LDA topics. From previous work, we know that LDA needs texts of a particular length in order to get reasonable probabilities to learn the topic model, but very short snippets (e.g., sentences) allow us to better identify the transitions in the text. Alternatively, we may use paragraphs as snippets, if we assume that no transition will occur within a paragraph.

Notes

[1] `https://www.nltk.org/`

[2] `https://spacy.io/`

[3] `https://scikit-learn.org`

References

Aurnhammer, C., Cuppen, I., van de Ven, I. & van Zaanen, M. (2019), 'Manual annotation of unsupervised models: Close and distant reading of politics on reddit.', *DHQ: Digital Humanities Quarterly* **13**(3).

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003*a*), 'Latent dirichlet allocation', *the Journal of machine Learning research* **3**, 993–1022.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003*b*), 'Latent dirichlet allocation', *J. Mach. Learn. Res.* **3**(null), 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. & Blei, D. (2009), Reading tea leaves: How humans interpret topic models, *in* Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams & A. Culotta, eds, 'Advances in Neural Information Processing Systems', Vol. 22, Curran Associates, Inc.

Collins, W. (1861), *Hide and Seek*, Sampson Low.

Genette, G., Lewin, J. E. & Culler, J. D. (1980), 'Narrative discourse : an essay in method', *Comparative Literature* **32**, 413.

Gupta, A., Srinivasan, P., Shi, J. & Davis, L. S. (2009), Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos, *in* '2009 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 2012–2019.

Harris, Z. (1954), 'Distributional structure', *Word* **10**(2-3), 146–162.

Huang, L. & Huant, L. (2013), Optimized event storyline generation based on mixture-event-aspect model, *in* 'Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing', pp. 726–735.

Jockers, M. L. & Mimno, D. (2013), 'Significant themes in 19th-century literature', *Poetics* **41**, 750–769.

Kodinariya, T. & Makwana, P. R. (2013), 'Review on determining number of cluster in k-means

clustering', *International Journal of Advance Research in Computer Science and Management Studies* **1**(6), 90–95.

Mill, J. S. (1861), *Utilitarianism*, Oxford University Press UK.

Moretti, F. (2013), *Distant Reading*, Verso, London.

Sbalchiero, S. & Eder, M. (2020), 'Topic modeling, long texts and the best number of topics. some problems and solutions', *Quality & Quantity* **54**, 1095–1108.

Syed, S. & Spruit, M. (2017), Full-text or abstract? examining topic coherence scores using latent dirichlet allocation, *in* '2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)', pp. 165–174.

Wallach, H. M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009), Evaluation methods for topic models, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ICML '09, Association for Computing Machinery, New York, NY, USA, p. 1105–1112.

Zhou, D., Xu, H., Dai, X.-Y. & He, Y. (2016), Unsupervised storyline extraction from news articles., *in* 'IJCAI', pp. 3014–3021.