Adapting a Topic Modelling Tool to the Task of Finding Recurring Themes in Folk Legends*

Maria Skeppstedt, Rickard Domeij, and Fredrik Skott

The Institute for Language and Folklore, Sweden firstname.lastname@isof.se

Abstract. A topic modelling tool, which was originally developed for performing text analysis on very short texts written in English, was adapted to the text genre of Swedish folk legends. The topic modelling tool was configured to use a word space model trained on a Swedish corpus, as well as a Swedish stop word list. The stop word list consisted of standard Swedish stop words, as well as 380 additional stop words that were tailored to the content of the corpus and therefore also included older spelling versions and grammatical forms of Swedish words. The adapted version of the tool was applied on a corpus consisting of around 10,000 Swedish folk legends, which resulted in the automatic extraction of 20 topics. Future versions of the tool will be extended with text summarisation functionality, in order to retain the text overview provided by the tool also when it is applied on longer folk legends.

Keywords: Folk Legends, Topic Modelling, Text Analysis.

1 Introduction

Topic modelling has been shown useful as a means of automatically identifying recurring information in large text collections. The technique has been applied on text collections belonging to a wide range of genres, e.g., news paper text [4], open-ended survey questions [2], Internet discussion forums [11], micro blogs [13] and student essays [7].

The topic modelling algorithm automatically analyses a text collection in search for topics, and then represents each detected topic by (i) a list of texts in which the topic is present, and (ii) a list of terms that frequently co-occur in these texts. There are a number of tools for visualising the output of the algorithm, for instance, visualisations that focus on relations between terms and topics, e.g., through presenting them in a matrix [5, 1]. We here instead focus on a tool, which builds on previous research that shows the potential in performing a manual analysis of the texts in which the detected topics are present. This tool, called Topics2Themes [12], therefore also visualises the extracted texts and their associations and provides support for performing a manual analysis of these texts.

Topics2Themes has so far mainly been developed for, and applied on, short, contemporary texts, e.g., micro blogs and posts in discussion forum threads. The collections have either contained English or Japanese texts. As a next step, we plan to apply the tool

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

^{*} Supported by the Swedish Research Council (project number 2017-00626).

on a text collection belonging to the genre of Swedish folk legends. There are previous approaches showing how topic modelling can be used for extracting information from this genre [6, 8]. However, since our collection of Swedish folk legends is very different from the texts on which Topics2Themes has previously been applied, adaptions of the tool have to be carried out. We will here discuss what adaptions are required in order to apply it on this new text genre.

2 The Topics2Themes tool

Figure 1 shows the output of the Topics2Themes¹ tool when the topic modelling algorithm has been applied on a collection of around 10,000 Swedish folk legends. The topic modelling algorithm Non-negative matrix factorisation [9] was used, and the algorithm was re-run ten times, only retaining topics stable enough to be included in all re-runs. This resulted in that 20, out of 25 requested, topics were retrieved by the tool.

The tool displays information in four different panels, as shown in the figure. The *Topics* panel, i.e., the second panel from the left, lists the topics that have been automatically extracted by the topic modelling algorithm. A topic element in the *Topics* panel is represented by a list element that contains the three terms most closely associated with the topic, e.g., "bonden – skogen – hörde" ("the farmer – the forrest – heard"). The *Terms* and *Texts* panels list the terms and texts, respectively, that are associated with the extracted topics. Term-topic and topic-text associations are visualised by lines that connect the list elements. Finally, to the *Themes* panel (only shown partly in the figure), the user can add recurring themes that are identified when the extracted texts are manually analysed. That is, while the first three panels stem from an automatic process, the content of the *Themes* panel is to be manually added by a user.

3 Adaptions to the genre of Swedish folk legends

The first adaption required was to configure the tool to use a Swedish stop word list in the stop word removal pre-processing step. We therefore configured the tool to use the standard Swedish stop word list included in NLTK [3], as well as 380 additional stop words tailored to the corpus. These 380 additional items mostly consisted of words belonging to modern day Swedish. However, the list also included frequent occurrences of words using older spelling versions, as well as old grammatical forms of words. For instance, "ock" ("and"), which is today spelled "och", and "blevo" ("became"), which is a plural verb form not used anymore, were added to the stop word list.

The second adaption consisted of configuring the tool to use a word space model trained on a Swedish corpus. We used a pre-trained Word2Vec continuous skipgram model² trained on the Swedish CoNLL17 corpus by the Language Technology Group at the University of Oslo. The word space model is used by the tool in a pre-processing step in which words are organised into groups based on semantic similarity. This is

¹ The code for the tool is available at: https://github.com/mariask2/topics2themes.

² The model has been trained on 3,010,472 tokens and is available at: http://vectors.nlpl.eu/repository/.



Fig. 1. Topics2Themes applied on a corpus of Swedish folk legends. The topic "kyrka – kyrkan – bygga" ("church – the church – build") is displayed with a blue background, which indicates that it has been selected by the user.

achieved through an automatic clustering of the word space vectors that represent the words included in the text collection [10]. The *Terms* panel in the figure shows examples of such term clusters. For instance, the first list element shows how the words "the church", "the church's" and "the churches" have been combined into one concept, and the eleventh list element shows a large word cluster of food-related terms.

4 Future work

In the current version of Topics2Themes, which is designed for shorter texts, the *Texts* panel is meant to give an easily scrollable overview of texts that have been extracted from the collection. This overview is, however, lost when the tool is applied on a collection that also contains longer texts, as fewer texts then can be shown simultaneously. The text presentation should therefore offer the functionality of showing shorter text summaries of retrieved texts, and let the user be able to toggle between viewing text summaries and the full version of the texts. Such a summarisation functionality will be implemented in a future version of the Topics2Themes tool.

Future work will also include an analysis of topics automatically extracted from the folk legends, as well as an evaluation of the usability of the tool, and of its usefulness for the task of finding recurring themes in texts belonging to the genre of Swedish folk legends.

References

- Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., Gleicher, M.: Serendip: Topic model-driven visual exploration of text corpora. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 173–182 (2014).
- Baumer, E.P.S., Mimno, D., Guha, S., Quan, E., Gay, G.K.: Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? Journal of the Association for Information Science and Technology (68-6), 1397–1410 (2017).
- Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA (2002).
- 4. Blei, D.M.: Topic modeling and digital humanities. Journal of Digital Humanities (2012).
- Chuang, J., Manning, C.D., Heer, J.: Termite: Visualization techniques for assessing textual topic models. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. pp. 74–77. ACM, New York, NY, USA (2012).
- Dagsson, T., Skott, F.: Digital cultural heritage a digital folklore archive. https://sweclarin.se/eng/ (2019).
- Ferrara, A., Montanelli, S., Petasis, G.: Unsupervised detection of argumentative units though topic modeling techniques. In: Proceedings of the 4th Workshop on Argument Mining. pp. 97–107. Association for Computational Linguistics, Copenhagen, Denmark (2017).
- 8. Karsdorp, F., den Bosch, A.V.: Identifying motifs in folktales using topic models. In: Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning (2013).
- 9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001).
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013).

- 11. Skeppstedt, M., Kerren, A., Stede, M.: Vaccine hesitancy in discussion forums: Computerassisted argument mining with topic models. In: Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. pp. 366–370. No. 247 in Studies in Health Technology and Informatics, IOS Press (2018).
- Skeppstedt, M., Kucher, K., Stede, M., Kerren, A.: Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In: Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources. pp. 9–16 (2018).
- Surian, D., Nguyen, D.Q., Kennedy, G., Johnson, M., Coiera, E., Dunn, A.G.: Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res (18-8), e232 (2016).