Snippets of Folk Legends: Adapting a Text Mining Tool to a Collection of 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 was adapted to requirements for a collection of Swedish folk legends. To offer an overview of a list of folk legend texts, which had been automatically extracted by the topic modelling tool, snippet text versions of the folk legends were displayed. The snippets were constructed from the full-text versions of the legends using the sentences most relevant to the topics extracted by the topic modelling algorithm. In addition, collection-adapted data was constructed for performing a pre-processing of the folk legend texts, before they were submitted to the topic modelling algorithm. This data consisted of a collection-adapted stop word list and word lists for improving the quality of clusters of semantically similar words.

Keywords: Folk Legends, Topic Modelling, Text Analysis.

1 Introduction

Topic modelling has been shown to be a useful text mining technique for 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 [3], open-ended survey questions [2], Internet discussion forums [12], micro blogs [14], student essays [5] and folk legends [6, 11].

The topic modelling algorithm automatically analyses a text collection in search for topics that often recur in the text. There are a number of topic modelling visualisation tools, for instance, visualisations that focus on relations between the automatically detected topics and the terms that represent these topics [1, 4]. We here instead use, and develop, a tool that builds on previous research which shows the potential in performing a manual analysis of the texts in which the topics detected are present. This tool, called Topics2Themes[1] [13], therefore also visualises the extracted texts and their associations and provides support for performing a manual analysis of these texts. The tool also provides a pre-processing step using a word2vec model [8], in which several words can be organised into groups based on semantic similarity, and treated as one concept. This is achieved through an automatic clustering of the word2vec vectors that represent the words included in the text collection.

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

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

Topics2Themes has previously been applied to posts in discussion forum threads [13], to micro blogs [10], and to research article paragraphs [9]. We recently applied the tool to a text collection consisting of around 10,000 Swedish folk legends [11]. The tool had previously been applied on English and Japanese texts, and to apply it on a Swedish text collection we needed to configure it to use a Swedish stop word list and a Swedish word2vec model. Apart from using standard Swedish stop words, we also added words typical to the text collection to the stop word list, e.g. words that use older spelling variants, and old grammatical forms.

Texts extracted by the topic modelling algorithm as most closely associated with the topics detected are displayed by Topics2Themes in a scrollable list. By scrolling through this text list, the user is meant to be given an overview of the content extracted. However, despite most of the folk legend texts being relatively short, many of them are longer than the very short types of texts for which the tool was originally developed. This has the effect that the overview, meant to be provided by the scrollable list, is lost. To achieve the same overview in those cases, we therefore added the functionality to – instead of displaying the full texts in the scrollable list – display a short snippet of the text. In addition to this technical adaption of the tool, we also (i) further adapted the stop word list to the folk legends collection by adding additional collection-specific words, and (ii) performed a manual correction of the automatically constructed concept clusters.

2 Topics2Themes Applied on the Folk Legend Collection

Topics2Themes was applied on a text collection containing around 10,000 records of folklore from the archive at the Institute for Language and Folklore in Gothenburg. The material mainly consists of folk legends, i.e. typically narratives about our perceptions of the world around us, about events in the past, and about everyday life manifestations of that, which we perceive as supernatural. The legends are usually relatively short and monoepisodic, have a fixed structure, and have been transmitted verbally from one generation to another.

Most of the records have been collected during the Interwar Period, through interviews with older people in the countryside in western Sweden. The records either contain (manual or HTR-based) transcriptions of handwritten text, or text obtained from printed material using OCR. Most of the material is written in Standard Swedish, although occasional dialectal expressions, as well as words written with older spelling variants, occur in the texts.

The Topics2Themes tool uses the topic modelling algorithm Non-negative matrix factorisation [7], and then represents each topic detected by the algorithm through, (i) a list of texts in which the topic is present, and (ii) a list of terms that frequently occur in these texts. We re-ran the topic modelling algorithm 50 times on our collection of folk legends, and only topics stable enough to be included in all re-runs were retained. This resulted in that 19, out of 25 requested, topics were retrieved by the tool.

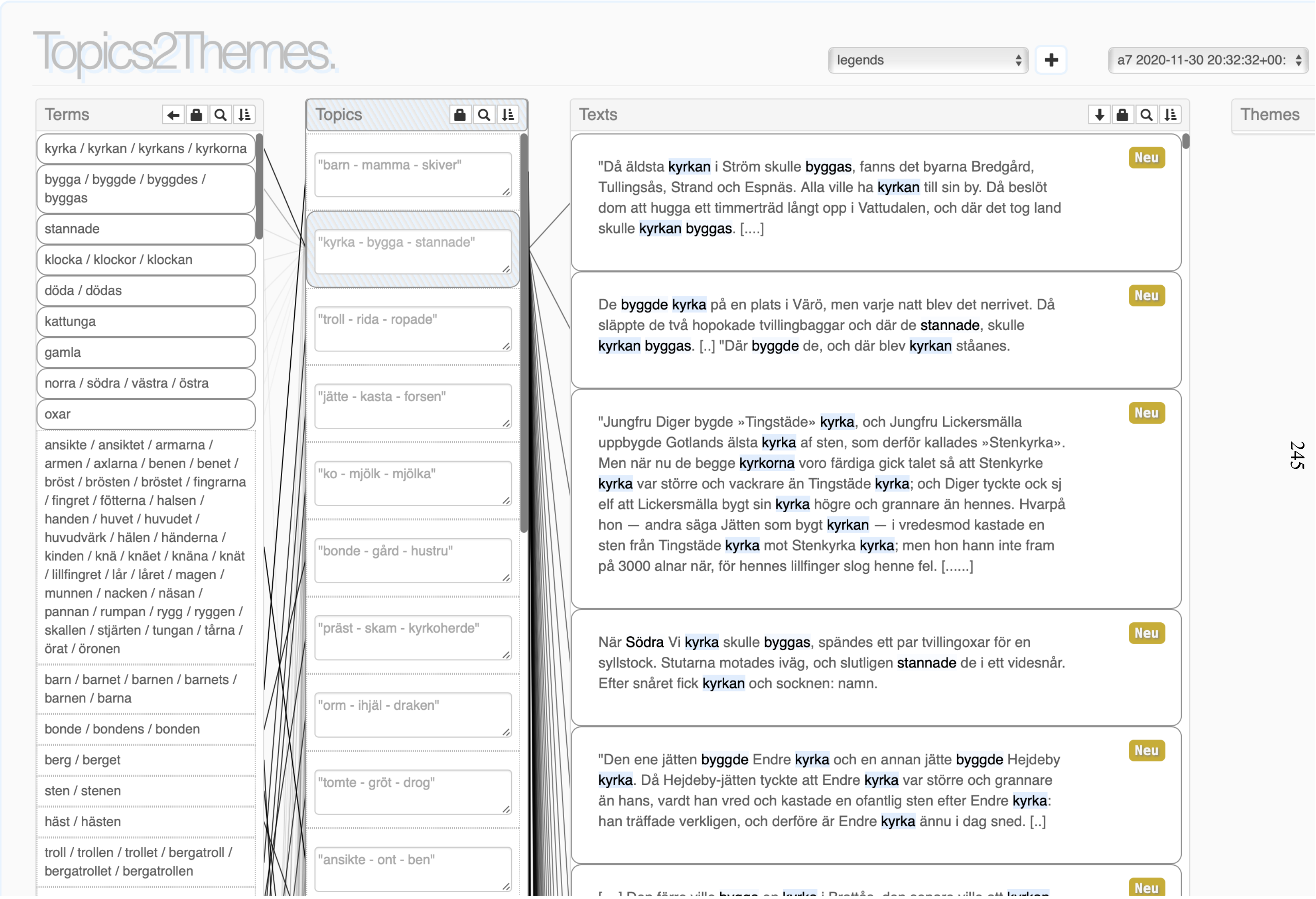
Figure 1 shows the graphical user interface of Topics2Themes, when applied on the folk legend collection. The tool displays information in four different panels. The *Topics* panel, i.e. the second panel from the left, lists the topics that have been automatically extracted by the topic modelling algorithm. 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. “kyrka – bygga – stannade” (“church – build – stayed”). The *Terms* and *Texts* panels list the terms and texts, respectively, that are associated with the extracted topics. We configured the tool to retrieve 10 terms and 40 texts for each topic. That is, for each topic detected, the *Terms* panel displays 10 terms, and the *Texts* panel displays 40 texts. 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.

The stop word list that was used for pre-processing the texts in the folk legends collection has been iteratively expanded. This process consists of first using Topics2Themes to construct a topic model, and thereafter of using the *Terms* panel to examine terms extracted (as well as the *Texts* panel to examine their textual contexts), in order to determine which terms are suitable as stop words. The process is then started over, using a list expanded with the newly detected stop words when constructing the new model. This iterative process has so far resulted in a stop word list consisting of 542 words.

For creating clusters of semantically similar words, we used a pre-trained word2vec continuous skipgram model[2] and the automatic clustering functionality of the Topics2Themes tool. We then used the functionality provided by the tool for manually correcting the automatically constructed clusters. This is carried out by providing the tool with two types of word lists, (i) a list of words that are to be removed from the automatic clusters (since the automatic process assigns them to a semantic cluster to which they do not belong), and (ii) a list of manually constructed word clusters. Our manually constructed clusters were often automatic clusters that we expanded with additional, semantically similar words, or automatic clusters that we divided into smaller, more semantically coherent groups of words. The construction of the word lists for improving the clustering is also an iterative process. We have so far detected 571 words to exclude from the automatic clustering, and constructed 118 manual clusters based on the automatic ones. A total of 369 concept clusters were used when pre-processing the texts before sending them to the topic modelling algorithm. Examples of concept clusters are shown in the *Terms* panel. For instance, the first element consists of different inflections of “church”, and the tenth element consists of a large cluster of words referring to body parts. The advantage of using semantic clustering instead of methods based on morphology – e.g. lemmatisation – is that semantically similar words can be conflated into one concept, regardless of their lexical instantiations.[3]

[2] The model has been trained on the Swedish CoNLL17 corpus – which contains 3,010,472 tokens – by the Language Technology Group at the University of Oslo.
The model is available at: http://vectors.nlpl.eu/repository/.

[3] The topic model configuration, as well as the manually constructed word list and clusters is available at: https://github.com/mariask2/swedish-folk-legends.



Terms

- kyrka / kyrkan / kyrkans / kyrkorna
- bygga / byggde / byggdes / byggas
- stannade
- klocka / klockor / klockan
- döda / dödas
- kattunga
- gamla
- norra / södra / västra / östra
- oxar
- ansikte / ansiktet / armarna / armen / axlarna / benen / benet / bröst / brösten / bröstet / fingrarna / fingret / fötterna / halsen / handen / huvet / huvudet / huvudvärk / hälen / händerna / kinden / knä / knäet / knäna / knät / lillfingret / lår / låret / magen / munnen / nacken / näsan / pannan / rumpan / rygg / ryggen / skallen / stjärten / tungan / tårna / örat / öronen
- barn / barnet / barnen / barnets / barnen / barna
- bonde / bondens / bonden
- berg / berget
- sten / stenen
- häst / hästen
- troll / trollen / trollet / bergatroll / bergatrollet / bergatrollen

Topics

- "barn - mamma - skiver"
- "kyrka - bygga - stannade"
- "troll - rida - ropade"
- "jätte - kasta - forsen"
- "ko - mjölk - mjölka"
- "bonde - gård - hustru"
- "präst - skam - kyrkoherde"
- "orm - ihjäl - draken"
- "tomte - gröt - drog"
- "ansikte - ont - ben"

Texts

"Då äldsta kyrkan i Ström skulle byggas, fanns det byarna Bredgård, Tullingsås, Strand och Espnäs. Alla ville ha kyrkan till sin by. Då beslöt dom att hugga ett timmerträd långt opp i Vattudalen, och där det tog land skulle kyrkan byggas. [....]

De byggde kyrka på en plats i Värö, men varje natt blev det nerrivet. Då släppte de två hopokade tvillingbaggar och där de stannade, skulle kyrkan byggas. [..] "Där byggde de, och där blev kyrkan ståanes.

"Jungfru Diger bygde »Tingstäde» kyrka, och Jungfru Lickersmälla uppbygde Gotlands älsta kyrka af sten, som derför kallades »Stenkyrka». Men när nu de begge kyrkorna voro färdiga gick talet så att Stenkyrke kyrka var större och vackrare än Tingstäde kyrka; och Diger tyckte ock sj elf att Lickersmälla bygt sin kyrka högre och grannare än hennes. Hvarpå hon — andra säga Jätten som bygt kyrkan — i vredesmod kastade en sten från Tingstäde kyrka mot Stenkyrka kyrka; men hon hann inte fram på 3000 alnar när, för hennes lillfinger slog henne fel. [......]

När Södra Vi kyrka skulle byggas, spändes ett par tvillingoxar för en syllstock. Stutarna motades iväg, och slutligen stannade de i ett videsnår. Efter snåret fick kyrkan och socknen: namn.

"Den ene jätten byggde Endre kyrka och en annan jätte byggde Hejdeby kyrka. Då Hejdeby-jätten tyckte att Endre kyrka var större och grannare än hans, vardt han vred och kastade en ofantlig sten efter Endre kyrka: han träffade verkligen, och derföre är Endre kyrka ännu i dag sned. [..]

Fig. 1. Topics2Themes applied on a collection of Swedish folk legends. The topic "kyrka – bygga – stannade" ("church – build – stayed") is displayed with a blue background in the center panel, which indicates that it has been selected by the user. The snippet versions of the texts most closely associated with the selected topic are shown to the right, while associated terms are indicated with a bold border to the left.

3 Producing the Snippet Text for the Extracted Sentences

The snippet text is produced by only retaining those sentences that contain terms that the topic modelling algorithm has extracted as typical to the topics. A sentence – or a sequence of sentences – that does not contain any of these terms is replaced by ellipsis, as shown in Figure 1. The Topics2Themes configuration file lets the user specify the maximum number of sentences to be included in the snippet text. This has the effect that only the first sentences in the original text that contain a term will be added to the text snippet, until the snippet has reached the maximum length specified. However, it is also made sure that the snippet text includes at least one context sentence for each one of the terms occurring in a text. Therefore, when more terms occur in the text than the maximum snippet length specified, the length of the snippet will instead be determined by the number of unique topic model-extracted terms occurring in the text.

The interface of Topics2Themes shows the snippet version of the text by default in the *Texts* panel. There are two methods by which the user can make the tool display the full-text version, (i) either by double-clicking on one specific list element to make the tool show the full-text version for this specific text, or (ii) by clicking on the arrow button in the header of the *Texts* panel to make the tool show the full-text versions for all texts.

4 Future Work

We have now carried out the technical developments required for adapting the Topics2Themes tool to our collection of folk legends. We have also provided the tool with collection-adapted data in the form of an adapted stop word list and word lists for improving the word2vec-based word clusters. As the next step, we will therefore use Topics2Themes for carrying out a computer-assisted analysis of the folk legend collection. That is, we will use the tool for identifying recurring themes in Swedish folk legends, i.e. themes that might (i) reflect already established classification systems, as well as (ii) provide new dimensions on how legends can be categorised in folkloristics.

References

1. 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)
2. 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)
3. Blei, D.M.: Topic modeling and digital humanities. Journal of Digital Humanities (2012)
4. 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)
5. 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)

6. 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)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556 – 562 (2001)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013), http://arxiv.org/abs/1301.3781, cite arxiv:1301.3781
9. Skeppstedt, M., Ahltorp, M., Eriksson, G., Domeij, R.: Annotating risk factor mentions in the COVID-19 Open Research Dataset. In: CLARIN2020 Book of Abstracts (2020)
10. Skeppstedt, M., Ahltorp, M., Kucher, K., Kerren, A., Rzepka, R., Araki, K.: Topic modelling applied to a second language: A language adaptation and tool evaluation study. In: Selected Papers from the CLARIN Annual Conference 2019. vol. 172:17, pp. 145–156. Linköping Electronic Conference Proceedings (2020)
11. Skeppstedt, M., Domeij, R., Skott, F.: Adapting a topic modelling tool to the task of finding recurring themes in folk legends. In: Proceedings of the Digital Humanities in the Nordic Countries. pp. 388–392. CEUR Workshop Proceedings (2020)
12. Skeppstedt, M., Kerren, A., Stede, M.: Vaccine hesitancy in discussion forums: Computer-assisted 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)
13. 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)
14. 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)