Master thesis proposals - Storytel and LiU

Måns Magnusson

21 oktober 2017

Below are master thesis proposals together with Storytel, one of Swedens largest audio book company (see storytel.se). These projects has been defined together with Storytel to cover issues that are of interest to them.

1 Thema-code classification of books

1.1 Background

When Storytel get books from external publishers, they get files (epub or mp3) as well as metadata for the books. These metadata files are in .xml format following a specific standard called ONIX[1]. Another standard is THEMA[2] which is a "global subject classification system for books". The classification is hiearchical with more general classes higher in the classification system. Today these codes are manually added to books by the publishers. This is of course both time consuming and could suffer from all sorts of flaws due to different human interpreting this standard differently. For some of the books this information is missing completely.

The purpose of this master thesis is to study if it is possible to automatically assign THEMA-codes to books given the raw text as input.

1.2 Thesis project

The thesis project will study different (hiearchical) text classification approaches to classify books with a very large number of classes. The student will propose classifications methods and evaluate different text classification approaches. As a baseline method an approach similar (but less complex) to the Kaggle hiearchical text classification winner will be used [9].

References

References of interest for the project include [9], [10] and [5].

Data and software

Storytel will be able to deliver text-data for books in English (21000), Swedish (12700), Danish (16800) and Norwegian (2700). Of these books some already have Thema codes. Any open source software may be used.

[1] See http://www.editeur.org/8/ONIX/
[2] See http://www.editeur.org/files/Thema/1.2/Thema_v1.2_en.pdf

2 Automatic identification and matching of litterature in different languages

2.1 Background

When Storytel get books from external publishers they get files (epub or mp3) as well as metadata-files. These metadatafiles are in .xml format following a the ONIX[3] standard. There's very limited support in the onix-standard to map (or know) when two books are actually the same literary work but just different editions/publications. Different language (translations) or different editions (just published by two different publishers, or in two different decades with some additional editing in between) makes it diffucult to handle large corpora computationally and be able to identify multiple translations or copies.

The purpose of this master thesis is to explore unsupervised methods to identify books that are translated and/or different editions. The idea would be to come up with methods, mainly using text data as input, to identify if two books are the "same" book, but in different languages.

2.2 Thesis project

The basic question is to find unsupervised methods to match individual books in different languages. The main question will be too study if polylingual topic models can be used to solve these problems as has been proposed in [6]. The basic polylingual topic model could be extended to capture more book specific structures (like character names) that could be used in book matching in a similar fashion to that proposed by [3].

An initial problem to use a polylingual topic model approach is that there need to be translated books to anchor topical structure between languages. To solve this we study the possibility to use either another parallel corpus, such as Wikipedia, or using Google translate for a subset of books.

The final unsupervised matching will eventually be evaluated by personel at Storytel of the approach seems promising.

References

References of interest for the project includes [7], [6] and [3]. Litterature in the field of text duplication identification may also be of interest such as [2] and [8].

Data and software

Storytel will be able to deliver text-data for (see below) books in English (21000), Swedish (12700), Danish (16800) and Norwegian (2700). No mapping exist today between any books (except for author).

Any open source software may be used, but the basic polya lingual topic model is already implemented in Mallet.

[3] See http://www.editeur.org/8/ONIX/

3 What make an (audio) book popular?

3.1 Background

Lately Storytel, similar to Netflix, has started to produce their own audiobook content. This makes it interesting for Storytel to study what aspects, such as author, literature quality, narrative and topics can explain the popularity of a given book, such as churn rates and the number of listeners.

3.2 Thesis project

The master thesis will try to identify what litterary qualities that may explain popularity and usage of Storytels books. This will result in both a focus on how to operationalize different literary qualities, such as narrative, as well as modeling and identifying what aspects of a book that correlates with popularity.

References

References of interest for the project are [1] and [4].

Data and software

Storytel will deliver metadata and text-data for a set of books together with some popularity score(s) and churn data. Preferable R will be used since a lot of suggested approaches has been presented in [4].

Additional information

This project may be co-supervised with a researcher in literary science, so the student should perferable be interested in literature as domain.

Referenser

[1] Jodie Archer and Matthew L Jockers. *The Bestseller Code: Anatomy of the Blockbuster Novel*. St. Martin's Press, 2016.

[2] Daniel Bär, Torsten Zesch, and Iryna Gurevych. Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012*, pages 167–184, 2012.

[3] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in neural information processing systems*, pages 241–248, 2007.

[4] Matthew Lee Jockers. *Text analysis with R for students of literature*. Springer, 2014.

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification, 2016.

[6] Kriste Krstovski and David A Smith. Online polylingual topic models for fast document translation detection. In *WMT@ ACL*, pages 252–261, 2013.

[7] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.

[8] Martin Potthast and Benno Stein. New issues in near-duplicate detection. *Data Analysis, Machine Learning and Applications*, pages 601–609, 2008.

[9] Antti Puurula, Jesse Read, and Albert Bifet. Kaggle lshtc4 winning solution. Technical report, 2014.

[10] Cao Ying and Duan run ying. Novel top-down methods for hierarchical text classification. *Procedia Engineering*, 24(Supplement C):329 – 334, 2011. International Conference on Advances in Engineering 2011.