Finding listening experiences in books

The Listening Experience Database Project (LED)[1] is an initiative aimed at collecting accounts of people's private experiences of listening to music [Barlow and Rowland (2017)]. Since 2012, the LED community explored a wide variety of sources, collecting over 10.000 unique experiences (see [Brown et al. (2014)] and [Adamou et al. (2014)]). The curatorial effort required to populate the database was significant and this result is a major achievement of the project.

Users start by exploring specific *sources of value*. These are books, for example, published by Internet Archive[2] or Google Books[3] and explored using either the search facility of the web portal or an application such as a PDF reader. The process starts from a source and moves to selecting a initial set of keywords. For example, *music*, sing*, song*, where consistently used, and then, elaborating from the retrieved material, expanded with more specific terms, in an iterative and exploratory process (illustrated in Figure 1).

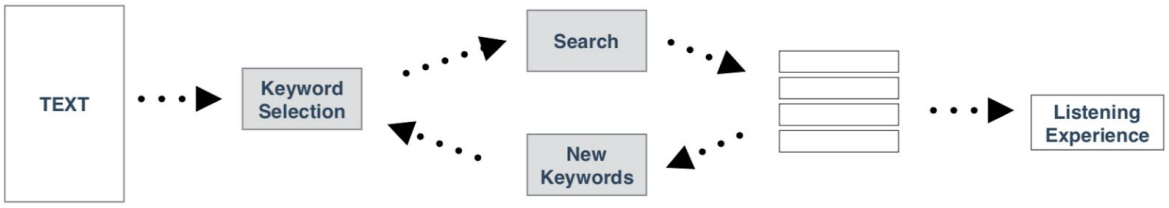

Figure 1. A sketch of the process for discovering a listening experience

This task requires effort, expertise, and is very time consuming. The process is not systematic or reproducible, therefore methodologically problematic. Ultimately, the decision to include a certain listening experience in the database is a curatorial choice. However, a systematic approach to finding candidate texts would boost the data acquisition process significantly.

In this work we report on the design of a system that supports curators in discovering listening experiences in texts. The task can be thought of as one of binary classification [Sokolova and Lapalme (2009)]. A book is segmented in a number of frames, and each one of them is evaluated by a *classifier* that assigns a positive or negative *label*. Clearly, the main question here is what approach should be taken in implementing such classifier. However, an equally important question is: how do we know that - whatever we do - would be *good enough*?

Gold standard

In natural language processing (NLP) the experiments aimed at evaluating and comparing the performance of different approaches are usually based on a reference corpus used as ground truth, or *gold standard* [Manning and Schütze (1999)]. We selected 500 positive samples from 17 books in LED. An equivalent number of negative samples were selected *from the same sources*. Figure 2 shows a pair of a positive and a negative sample. Notably,

[1] LED Project,http://www.listening-experience.org/. Accessed 16 October 2018.
[2] Internet Archive: https://archive.org/, accessed 22 October 2018.
[3] Google Books:https://books.google.co.uk/, accessed 22 October 2018.

the negative sample includes a number of terms referring to music but the text itself does not report a listening event.

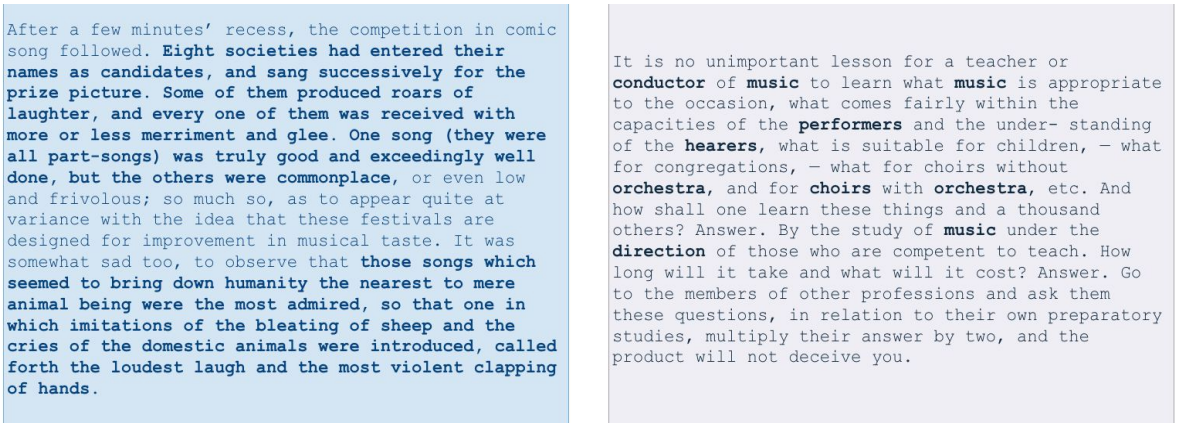

Figure 2. Comparison of a positive sample (on the left) and a negative one.

Competing methods

We developed and compared three fundamental approaches. All the methods treat a text as vector of POS-tagged keywords generated with standard NLP techniques (see Figure 3).

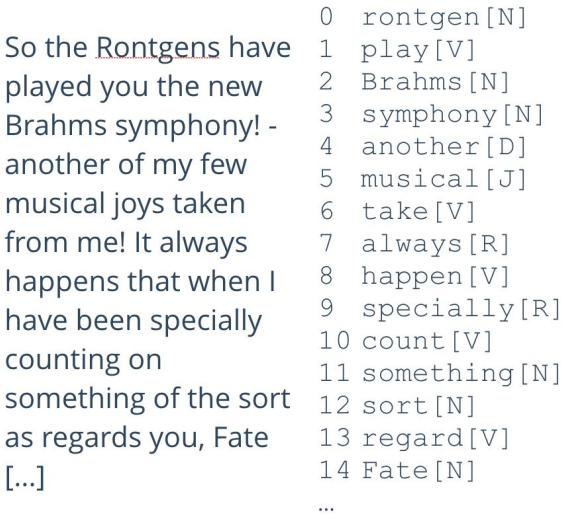

Figure 3. Example of a text represented as a vector of POS-tagged words.

Musical Forest. The first approach is based on the idea that we can learn the features of listening experiences adopting a typical machine learning workflow and relying on the LED Database. We chose a Random Forest algorithm [Ho (1995)] as implemented by Apache Spark.[4]

Musical Gut. An alternative way of approaching the problem is to use a dictionary of words in the musical domain. To this aim, we can apply *statistical* NLP techniques. Project Gutenberg publishes approximately 50 thousands english books in the public domain.[5]

[4] Apache Spark: https://spark.apache.org/ Accessed 15th November 2018.
[5] Project Gutenberg: https://www.gutenberg.org/ Accessed 15th November 2018.

Fortunately, it also includes a Music shelf. We combined all the words occurring in books in the Music shelf and computed the average TF/IDF value to obtain a dictionary that we used to estimate the relatedness of a text to the music domain..

Musical Predictions. A neural network (NN) can be trained to *predict* words that can appear in the same context [Mikolov (2013)]. This approach will generate so-called word *embeddings*. From these we can extract a dictionary of terms related to music.

Experimental evaluation

As a classification system, the performance can be measured as *accuracy*.[6] As a Information Retrieval system, the performance is the capacity to return positive results and it is calculated as *F-measure*.[7] Our experiments employed three annotators developed on top of the Stanford NLP library[8] [Manning (2014)] and our gold standard. Results are summarized by Table 1.

The most performing annotator is the one based on word embeddings. We used it to develop a novel system to support curators in discovering traces of listening experiences in texts.

Method	**Precision**	**Recall**	**F1**	**Accuracy**
Forest	0.52	0.99	0.69	0.55
Gut	0.72	0.95	0.82	0.79
Predictions	0.82	0.91	**0.86**	**0.85**

Table 1. Comparison between the three methods.

Discovery of listening experiences with FindLEr

FindLEr[9] supports users in the discovery of traces of listening experiences in texts (see Figure 6). The curator can provide a source book as URI or file and obtain a annotated version of the text where paragraphs mentioning potential listening experiences are highlighted. The user can browse the results as a list or inline with the original text. The system allows for additional tuning by offering a *skepticism* handle. Increasing the value will make the system more selective and return less results. On the contrary, reducing the skepticism will increase the number of matches. The user can notify the system about the quality of each result contributing to enrich the set of positive and negative examples to be used in the future for further improving on the underlying method.

[6] Accuracy: https://en.wikipedia.org/w/index.php?title=Accuracy_and_precision&oldid=886339200
[7] F-measure: https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=886289077
[8] Stanford Core NLP: https://stanfordnlp.github.io/CoreNLP/ Accessed 15th November 2018.
[9] FindLEr: https://led.kmi.open.ac.uk/discovery . Accessed 15th November 2018.

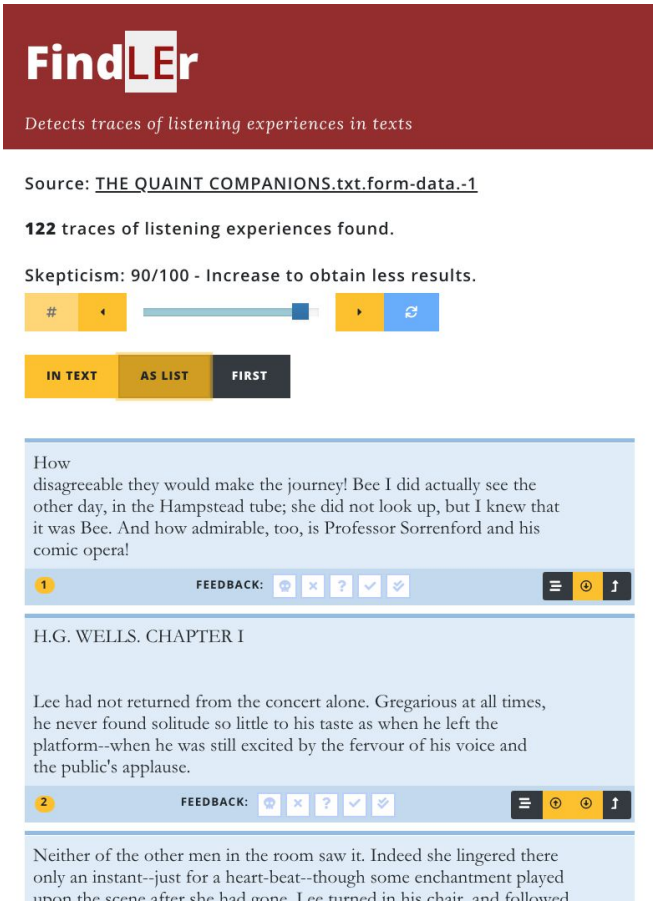

Figure 4. The FindLEr web application.

Discussion and perspectives

The problem of keyword expansions and concept-based search is common in digital humanities research [Cheema (2016)] [Osadetz (2018)] and the overcoming of keyword-based approaches a recurring theme in computer science (e.g. [Finkelstein (2002)] and [Giunchiglia (2009)]). However, this is the first attempt of applying state of the art computational methods for finding listening experiences in texts. Applying the system to repositories instead of books would be an interesting engineering challenge, for example selecting relevant sources from a registry of datasets and repositories such as *musoW* [Daquino et al. (2017)].[10]

References

Adamou, A., M. d'Aquin, H. Barlow, and S. Brown (2014). LED: curated and crowd-sourced linked data on music listening experiences. Proceedings of the ISWC 2014 Posters & Demonstrations Track, 93–96.

Barlow, Helen and Rowland, David (Ed.) (2017). Listening to music: people, practices and experiences. The Open University.

[10] Musical Data on the Web: http://musow.kmi.open.ac.uk/

Brown, S., A. Adamou, H. Barlow, and M. d'Aquin (2014). Building listening experience linked data through crowd-sourcing and reuse of library data. In Proceedings of the 1st International Workshop on Digital Libraries for Musicology, pp. 1–8. ACM.

Cheema, Muhammad Faisal, et al. (2016) "A Directed Concept Search Environment to Visually Explore Texts Related to User-defined Concept Models." VISIGRAPP (2: IVAPP). 2016.

Daquino, M., Daga, E., d'Aquin, M., Gangemi, A., Holland, S., Laney, R., ... & Mulholland, P. (2017). Characterizing the Landscape of Musical Data on the Web: State of the art and challenges. In Proceedings of the 2nd Workshop of Digital Humanities in the Semantic Web (WHiSe). Ceur-WS.

Finkelstein, Lev, et al. (2012) "Placing search in context: The concept revisited." *ACM Transactions on information systems* 20.1 (2002): 116-131.

Giunchiglia, Fausto, Uladzimir Kharkevich, and Ilya Zaihrayeu. (2009) "Concept search." *European Semantic Web Conference*. Springer, Berlin, Heidelberg, 2009.

Ho, Tin Kam. (1995) "Random decision forests." Document analysis and recognition, 1995., proceedings of the third international conference on. Vol. 1. IEEE, 1995.

Manning, Christopher D. and Hinrich Schütze. (1999) *Foundations of statistical natural language processing*. MIT press, 1999.

Mikolov, Tomas, et al. (2013) "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Osadetz, Stephen, et al. (2018) "Searching for Concepts in Large Text Corpora: The Case of Principles in the Enlightenment." Digital Humanities Conference (DH). 2018.

Sokolova, M. and G. Lapalme (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4), 427–437.