Enriching Digital Collections Using Tools for Text Mining, Indexing and Visualization

Riza Batista-Navarro[1](✉), Axel J. Soto[1], Nhung T.H. Nguyen[1], William Ulate[2], and Sophia Ananiadou[1]

[1] School of Computer Science, University of Manchester, Manchester, UK
{friza.batista,axel.soto,nhung.nguyen, sophia.ananiadoug}@manchester.ac.uk
[2] Missouri Botanical Garden, St. Louis, MO, USA
william.ulate@mobot.org

1 Aims, Scope and Learning Objectives

The tutorial demonstrated a suite of tools for text mining, semantic indexing and visualization that facilitated enhanced searching and exploration of digital collections. Specifically, we aimed to provide:

- an introduction to modular text mining and indexing workflows developed using the Argo platform (http://argo.nactem.ac.uk);
- an overview of the Elasticsearch indexing engine and the Kibana visualization platform;
- the know-how on building and visualizing semantic indexes over digital collections without any programming effort.

The tutorial will cover the end-to-end automatic generation and visualization of a semantically enabled search index over digital collections. By the end of the tutorial, the audience will have gained knowledge on:

- exemplar digital collections (e.g., the Biodiversity Heritage Library, British Medical Journal) enhanced with text-mined semantic metadata and visualization tools;
- information extraction methods for generating semantic metadata over textual collections;
- employing Argo to construct text mining workflows that generate Elastic-search indexes for searching over digital collections;
- using the Kibana platform to generate dashboards and visually explore digital collections indexed with Elasticsearch.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 665–666, 2017.
DOI: 10.1007/978-3-319-67008-9

2 Relevance to TPDL 2017 and Significance to the Field

With the vast amounts of heterogeneous data that many digital libraries hold, finding information relevant to users has become a challenge. One of the most complex types of data is text written in natural language, whose unstructured and ambiguous nature poses a barrier to the accessibility and discovery of information. Furthermore, the volume of available data makes the exploration and discovery of meaningful content difficult. This can be alleviated by means of a combination of semantic indexing and interactive visualization. Firstly, documents can be indexed with semantic metadata, e.g., by tagging them with terms that indicate their "aboutness". As manually indexing these documents is impractical, automatic tools capable of generating semantic metadata and building search indexes have become attractive solutions. Secondly, users of digital libraries need to be provided with the necessary tools to explore collections and be able to quickly answer analytical questions based on the data. Information visualization, therefore, represents a valuable asset as it aims at showing summarized information in an intuitive manner.

In this tutorial, we aimed to demonstrate how digital library developers and managers (who do not necessarily have the expertise on natural language processing, text mining and visualization) can make their digital collections easier to search and explore. To this end, we showed how Argo can facilitate the development of their own customized, modular workflows for automatic semantic metadata generation and search index construction. Moreover, we showed how the Kibana platform can be used to slice and dice different views of the data and facilitate their visual exploration.

In this way, the tutorial provided digital library practitioners with the necessary technical know-how on building and visualizing semantic search indexes without any programming effort. We believe that this in turn will allow various digital libraries to build search systems that enable users to find and discover information of interest in a more scalable and efficient manner.

3 Target Audience

This tutorial did not require any prerequisite knowledge and the concepts which will be presented were at the introductory level. Although its aim was to enable the audience to build a technical artifact, no knowledge of programming was required, owing to the Argo platform's graphical interface for workflow construction as well as Kibana's ready-to-use visualizations based on Elasticsearch queries. We wished to reach out to attendees who were developers and managers of digital libraries interested in enhancing their systems with capabilities that facilitate semantic searching and visualization over digital collections.