Science of Computer Programming 121 (2016) 128–152

Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico

TextFlows: A visual programming platform for text mining and natural language processing

Matic Perovšek [a,b,*], Janez Kranjc [a,b], Tomaž Erjavec [a,b], Bojan Cestnik [a,c], Nada Lavrač [a,b,d]

[a] Jožef Stefan Institute, Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[c] Temida d.o.o., Ljubljana, Slovenia
[d] University of Nova Gorica, Nova Gorica, Slovenia

ARTICLE INFO

Article history:
Received 14 March 2015
Received in revised form 28 December 2015
Accepted 5 January 2016
Available online 14 January 2016

Keywords:
Text mining
Natural language processing
Web platform
Workflows
Human-computer interaction

ABSTRACT

Text mining and natural language processing are fast growing areas of research, with numerous applications in business, science and creative industries. This paper presents TextFlows, a web-based text mining and natural language processing platform supporting workflow construction, sharing and execution. The platform enables visual construction of text mining workflows through a web browser, and the execution of the constructed workflows on a processing cloud. This makes TextFlows an adaptable infrastructure for the construction and sharing of text processing workflows, which can be reused in various applications. The paper presents the implemented text mining and language processing modules, and describes some precomposed workflows. Their features are demonstrated on three use cases: comparison of document classifiers and of different part-of-speech taggers on a text categorization problem, and outlier detection in document corpora.

1. Introduction

Text mining [9] is a research area that deals with the construction of models and patterns from text resources, aiming at solving tasks such as text categorization and clustering, taxonomy construction, and sentiment analysis. This research area, also known as text data mining or text analytics, is usually considered as a subfield of data mining (DM) research [12], but can be viewed also more generally as a multidisciplinary field drawing its techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), information extraction (IE) and knowledge management.

From a procedural point of view, text mining processes typically follow the CRISP-DM reference process model for data mining [7], which proposes six phases when working on a DM project: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Text mining can be distinguished from general data mining by special procedures applied in the data preparation phase, where unstructured or poorly structured text needs to be converted into organized data, structured as a table of instances (rows) described by attributes (columns). In the modeling phase, such a table of instances can be used by the standard or slightly adapted data mining algorithms to uncover interesting

* Corresponding author.
E-mail addresses: matic.perovsek@ijs.si (M. Perovšek), janez.kranjc@ijs.si (J. Kranjc), tomaz.erjavec@ijs.si (T. Erjavec), bojan.cestnik@temida.si (B. Cestnik), nada.lavrac@ijs.si (N. Lavrač).

http://dx.doi.org/10.1016/j.scico.2016.01.001
0167-6423/

information hidden in the data. Two typical approaches are using clustering algorithms to find groups of similar instances or classification rule learning algorithms to categorize new instances.

The TextFlows platform described in this paper is a new open source, web-based text mining platform that supports the design and composition of scientific procedures implemented as executable workflows. As a fork of ClowdFlows [28], TextFlows has inherited its service-oriented architecture that allows the user to utilize arbitrary web services as workflow components. TextFlows is oriented towards text analytics and offers a number of algorithms for text mining and natural language processing. The platform is implemented as a cloud-based web application and attempts to overcome various deficiencies of similar text analytics platforms, providing novel features that should be beneficial to the text mining community. In contrast to existing text analytics workflow management systems, the developed platform is the only one with all the following properties. It is simple (i.e., enables visual programming, is web-based and requires no installation), enables workflow sharing and reuse, and is open source. Moreover, the platform enables combining workflow components (called "widgets") from different contexts (e.g., using clustering in relational data mining) and from different software sources (e.g., building ensemble classifiers from different libraries). To do so, it provides a unified input-output representation, which enables interoperability between widget libraries through automated data type conversion. It uses a common text representation structure and advocates the usage of 'hubs' for algorithm execution.

TextFlows is publicly available at http://textflows.org, while its source code is available at https://github.com/xflows/textflows under the MIT License. Detailed installation instructions are provided with the source code. After setting up a local TextFlows instance, advanced users can also implement and test their own algorithms. Improvements to the code can also be pushed to the main Git code repository via pull requests. The committed changes are reviewed by the TextFlows core team and merged into the master branch.

TextFlows is a web application which can be accessed and controlled from anywhere while the processing is performed in a cloud of computing nodes. TextFlows differs from most comparable text mining platforms in that it resides on a server (or cluster of machines) while its graphical user interface for workflow construction is served as a web application through a web browser. The distinguishing feature is the ease of sharing and publicizing the constructed workflows, together with an ever growing roster of reusable workflow components and entire workflows. As not only widgets and workflows, but also data and results can be made public by the author, TextFlows can serve as an easy-to-access integration platform both for various text mining workflows but also for experiment replicability. Each public workflow is assigned a unique URL that can be accessed by anyone to either repeat the experiment, or to use the workflow as a template to design another, similar, workflow.

Workflow components (widgets) in TextFlows are organized into packages which allows for easier distributed development. The TextFlows packages implement several text mining algorithms from LATINO[1] [11], NLTK[2] [3] and scikit-learn[3] [32] libraries. Moreover, TextFlows is easily extensible by adding new packages and workflow components. Workflow components of several types allow graphical user interaction during run-time and visualization of results by implementing views in JavaScript, HTML or any other format that can be rendered in a web browser (e.g., Flash, Java Applet).

The rest of the paper is structured as follows. Section 2 presents the technical background and implementation details of the TextFlows platform, along with its key text mining components. The architecture of the system is presented in detail along with specific data structures that allow efficient text mining in a workflow environment. The concept of workflows, their implementation, execution and sharing are presented in Section 3, while Section 4 describes the widget repository and the implemented modules of the platform. The advanced features of TextFlows are demonstrated in Section 5 on three use cases: a comparison of document classifiers on a classification problem, a comparison of part-of-speech taggers on a text categorization (classification) problem, and outlier detection in document corpora. Section 6 presents the related work, where we describe comparable text mining platforms together with their similarities and differences compared to the TextFlows platform. Section 7 concludes the paper by presenting a summary and some directions for further work.

2. The TextFlows platform

This section presents the TextFlows platform, together with its architecture and main components of the system. We also introduce the graphical user interface and describe the concept of workflows. Like its predecessor data mining platform ClowdFlows [28], TextFlows can also be accessed and controlled from a browser, while the processing is performed on a cloud of computing nodes. In this section we explain the relationship between TextFlows and ClowdFlows, present the architecture of the TextFlows platform and describe the key text mining concepts of TextFlows in more detail.

2.1. Platform architecture

In software engineering, terms *front-end* and *back-end* are used to distinguish the separation between a presentation layer (the client side) and a data access layer (the server side), respectively. Fig. 1 shows the TextFlows architecture, which

[1] LATINO (Link Analysis and Text Mining Toolbox) is open-source—mostly under the LGPL license—and is available at http://source.ijs.si/mgrcar/latino.
[2] http://www.nltk.org.
[3] http://scikit-learn.org.