
Aprosil nime verec andi v vum, quis nulentor aut porpum sim ut diorem vendunt d um aut autolit eris d.

To process texts with the KNIME Text Processing extension, exact quias minvemate areunt aliquas derunt a number of steps are required. Faccabor rem re natur, siti volonp quisit aut a dii que sitas quid que dolor a voluptato forem dawn eculeplo dolorei sam quis sellitkareeque nessinm quatam ex es et atas veratur. Conistanti biecist runqqui seq Percii leustrumque lab intur, id eos testias pashit lanti te delectatem voqesai comam rectur, r duvas maximior rest, consequi inus. Altur? Gounduten ad eius veligendion et aut verum alitibus et militia volut inum faccat illum uuptatexti tem sumquat emporac epltur* ullam veligendion et aut verum curs maxim eosct, occass est autem aut a dolupta situs et.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td>10</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>11</td>
</tr>
<tr>
<td>Chapter 1. Introduction</td>
<td>12</td>
</tr>
<tr>
<td>1.1. Why Text Mining?</td>
<td>12</td>
</tr>
<tr>
<td>1.2. Install the KNIME Text Processing Extension</td>
<td>12</td>
</tr>
<tr>
<td>1.3. Data Types for Text Processing</td>
<td>14</td>
</tr>
<tr>
<td>1.4. The Text Mining Process</td>
<td>15</td>
</tr>
<tr>
<td>1.5. Goals and Organization of this Book</td>
<td>17</td>
</tr>
<tr>
<td>Chapter 2. Access Data</td>
<td>19</td>
</tr>
<tr>
<td>2.1. Introduction</td>
<td>19</td>
</tr>
<tr>
<td>2.2. Read Text Data and Convert to Document</td>
<td>20</td>
</tr>
<tr>
<td>Strings To Document</td>
<td>22</td>
</tr>
<tr>
<td>2.3. The Tika Integration</td>
<td>23</td>
</tr>
<tr>
<td>Tika Parser</td>
<td>24</td>
</tr>
<tr>
<td>2.4. Access Data from the Web</td>
<td>27</td>
</tr>
<tr>
<td>2.4.1. RSS Feeds</td>
<td>27</td>
</tr>
<tr>
<td>RSS Feed Reader</td>
<td>27</td>
</tr>
<tr>
<td>2.4.2. Web Crawling</td>
<td>28</td>
</tr>
<tr>
<td>HttpRetriever</td>
<td>29</td>
</tr>
<tr>
<td>HtmlParser</td>
<td>30</td>
</tr>
<tr>
<td>XPath</td>
<td>31</td>
</tr>
<tr>
<td>Content Extractor</td>
<td>32</td>
</tr>
<tr>
<td>2.5. Social Media Channels</td>
<td>33</td>
</tr>
<tr>
<td>2.5.1. Twitter</td>
<td>33</td>
</tr>
</tbody>
</table>
Chapter 3. Text Processing

3.1. What is Text Processing? ............................................................... 46

3.2. Enrichment: Taggers .................................................................. 46

3.2.1. Part-Of-Speech Taggers .......................................................... 48

   POS Tagger .......................................................................................... 48
   Stanford Tagger .................................................................................. 52

3.2.2. Domain Taggers ..................................................................... 53

   OpenNLP NE Tagger ......................................................................... 54
   Abner Tagger .................................................................................... 56
   OSCAR Tagger .................................................................................. 56

3.2.3. Custom Taggers ...................................................................... 57

   Dictionary Tagger ............................................................................... 57
   Wildcard Tagger ................................................................................ 59

3.3. Filtering ..................................................................................... 60
4.1. Chapter 4. Frequencies and Vectors

3.4. Stemming and Lemmatization
- Porter Stemmer .............................................. 66
- Snowball Stemmer ........................................... 67
- Stanford Lemmatizer ........................................ 68

3.5. Bag of Words .................................................. 69
- Bag Of Words Creator ......................................... 70

3.6. Helper Nodes .................................................. 72
- Document Data Extractor ..................................... 73
- Sentence Extractor ............................................ 74
- Meta Info Inserter ............................................ 74
- Tag to String .................................................... 75

3.5. Exercises ....................................................... 76
- Exercise 1 ......................................................... 76
- Exercise 2 ......................................................... 77
- Exercise 3 ......................................................... 77

Chapter 4. Frequencies and Vectors ........................................ 79

4.1. From Words to Numbers ....................................... 79

4.2. Word Frequencies ............................................. 79
- TF ................................................................. 81
- IDF ............................................................... 83

6
Foreword

From scientific papers, over Wikipedia articles, patents, tweets, to medical case reports, and product reviews, textual data is generated and stored in various areas, to document, educate, tell, influence, or simply to entertain. Not only the amount of textual data is growing massively every year, also the areas in which text is generated and can be mined are increasing.

Due to the complexity of human natural language and the unstructured and sequential nature of the data, it is especially complex to mine and analyze text. In order to handle this complexity, specific methods have been invented in the fields of text mining and natural language processing. Whereas pure text mining is focusing on the extraction of structured knowledge and information from text, natural language processing is approaching the problem of the understanding of natural language.

Many of these methods and algorithms have been implemented in a variety of tools and platforms. For example, important open source libraries are Stanford NLP and Apache OpenNLP as well as packages in R and Python. Both of them, Stanford NLP and Apache OpenNLP, are integrated in the KNIME Text Processing extension. Due to the visual programming paradigm of KNIME, the Text Processing extension enables also non-programmers and non-scripters, not only to use those libraries, but also to easily combine them with a variety of other functionalities.

Still, text mining is not an easy task, even with the right tool. Text processing functionality needs to be well understood and correctly used, before applying them. This is why this book will prove to be extremely helpful. Also the timing for the book release is perfect: The KNIME Text Processing extension was moved out of KNIME Labs* recently, with the release of the KNIME Analytics Platform version 3.5.

Rosaria and Vincenzo have done an outstanding job writing this truly comprehensive book describing the application of text mining and text processing techniques via the KNIME Text Processing extension in combination with other KNIME Analytics Platform data science resources.

Kilian Thiel

* KNIME Labs category in KNIME Analytics Platform is dedicated to advanced and not yet fully established data science techniques.
When writing a book, it is impossible not to ask and learn from a few people. That was the case for this book as well. So, here it is our chance to thank all those people who taught us more about text mining, who provided us with some level of technical support, who gave us interesting ideas, and, in general, who have stood us through these last few months. Here they are.

First of all, we would like to thank Kilian Thiel for explaining how a few mysterious nodes are working. Kilian, by the way, was the developer zero of the KNIME Text Mining extension.

We would like to thank Heather Fyson for correcting our writing and, especially, for anglicizing our English from the strong Italian influences.

Frank Vial is responsible for exactly four words in this book: the title.

Finally, a word of thanks to Kathrin Melcher and Adrian Nembach who provided precious help for the neural network and deep learning part.
Chapter 1. Introduction

1.1. Why Text Mining?
We often hear that we are in the age of data [1], that data may become more important than software [2], or that data is the new oil [3], but much of these data are actually texts. Blog posts, forum posts, comments, feedbacks, tweets, social media, reviews, descriptions, web pages, and even books are often available, waiting to be analyzed. This is exactly the domain of text mining.

**KNIME Analytics Platform** offers a text processing extension, fully covering your needs in terms of text analytics. This extension relies on two specific data objects: the Document and the Term.

A Document object is not just text, but it also includes the text title, author, source, and other information. Similarly a Term is not just a word, but it includes additional information, such as its grammar role or its reference entity.

The **KNIME Text Processing extension** includes nodes to read and write Documents from and to a variety of text formats; to add word information to Terms; to clean up sentences from spurious characters and meaningless words; to transform a text into a numerical data table; to calculate all required word statistics; and finally to explore topics and sentiment.

The goal of this book is to explore together all steps necessary and possible to pass from a set of texts to a set of topics or from a set of texts to their in between the lines sentiments.

1.2. Install the KNIME Text Processing Extension
The KNIME Text Processing extension, like all KNIME extensions, can be installed within the KNIME Analytics Platform from the top menu items:

- File -> Install KNIME Extensions …

Or

- Help -> Install New Software …

Both menu items open to the “Install” window.
• In the text box labelled “Work with:” connect to the KNIME Analytics Platform Update Site (i.e. ‘http://update.knime.com/analytics-platform/3.5’ for KNIME Analytics version 3.5);
• Expand item “KNIME & Extensions” and select extension “KNIME Text Processing” and the language packs you wish to use;
• Click “Next” and follow the installation instructions.

If installation has been successful, you should end up with a category Other Data Types/Text Processing in the Node Repository panel. No additional installation is required, besides downloading occasional dictionary files for specific languages. Usually such dictionary files can be found at academic linguistic departments, like for example at the WordNet site.

![Figure 1.2. Settings for the Text Processing extension in the Preferences window](image)

After the installation of the KNIME Text Processing extension, you can set a few general preferences for the Text Processing nodes.
Under Preferences -> KNIME -> Text Processing, you can set the tokenizer properties. Here you can also set how to store text data cells and, in case of file based storage, the chunk size; that is the number of Documents to store in a single file. Finally, you can define the list of search engines appearing in the Document view, allowing the search for meaning or synonyms.

1.3. Data Types for Text Processing

Nodes in the KNIME Text Processing extension relies on two new types of data: Document and Term.

A raw text becomes a Document when additional metadata, such as title, author(s), source, and class, are added to the original text. Text in a Document gets tokenized following one of the many tokenization algorithms available for different languages. Document tokenization produces a hierarchical structure of the text items: sections, paragraphs, sentences, and words. Words are often referred to as tokens. Below you can see an example of the hierarchy produced by the tokenization process applied to an email.

Similarly to the Document object, a token becomes a Term with the addition of related metadata, and specifically tags. Tags describe sentiment, part of speech, city (if any), person name (if any), etc... covered by the word in the Term. Below you can see a few Term examples from the sentence “I love Sevilla”.

Term “I” includes token (word) “I” and it’s Part Of Speech = “Pronoun”.

Term “love” includes token (word) “love”, Part Of Speech = “Verb”, and Sentiment = “Positive”.

Figure 1.3. Tokenization items in an email Document
Term “Sevilla” includes token (word) “Sevilla”, Part Of Speech = “Noun”, and Named Entity = “City”.

Figure 1.4. Term structure from the sentence “I love Sevilla”.

1.4. The Text Mining Process

The whole goal of text data preparation is to convert the text into numbers, as to be able to analyze it with all available statistical and machine learning techniques.

The process always starts with text reading, whatever the text format is.

After that, we transform the simple text String into a more complex Document object. For this transformation, a tokenization operation is required. Tokenization algorithms identify and label parts of the input texts as sections, paragraphs, sentences, and terms. Once all those text parts have been detected, labelled, and stored, the Document object is born.

After defining the hierarchical structure of the Document, it is possible to attach specific tags to some terms, such as grammar roles (Part Of Speech, POS), sentiment, city names, general entity names, dictionary specific tags, and so on. This tagging operation is named enrichment, since it enriches the information content of the Term.
Now that we have tokenized the text down to Terms and that we have included extra information in some of the Terms, if not all, we can proceed with more aggressive **clean up**. The main goal of the cleanup phase is to get rid of all those words carrying too little information. For example, prepositions and conjunctions are usually associated with grammar rules, rather than with semantic meaning. These words can be removed using:

- A tag filter, if a POS tagging operation has been previously applied;
- A filter for short words, i.e. shorter than N characters;
- A filter for stop words, specifically listed in a dictionary file.

Numbers could also be removed as well as punctuation signs. Other ad hoc cleaning procedures could also help to make the Document content more compact. Cleanup procedures usually go together with other generic pre-processing steps.

A classic pre-processing step consists of **stemming**, i.e. of extracting the word stem. For example, the words “promising” and “promise” carry the same meaning in two different grammar forms. With a stemming operation, both words would be reduced to their stem “promis[]”. The stemming operation makes the word semantic independent of the grammar form.

Now we are ready to collect the remaining words in a **bag of words** and to assign a frequency based score to each one of them. If the words in the bag of words are too many, even after the text cleaning, we could consider the option of summarizing a Document through a set of **keywords**. In this case, all words receive a score, quantifying their summary power, and only the top n words are kept: the n keywords. Words/keywords with their corresponding score pass then to the next phase: **Transformation**.

![Figure 1.5. The many phases of a Text Analytics process](image)

**Transformation** covers **encoding** and **embedding**. Here the Document moves from being represented by a set of words to being represented by a set of numbers. When using encoding we refer to the presence (1) / absence (0) of a word in a Document text: 1 if the word is present, 0 if it is absent. We then define a matrix where each word gets a dedicated column and each Document is represented by a sequence of 0s and 1s, depending on the presence/absence of each column word in the Document. Instead of 1s the frequency based score of the word could also be used. Embedding is another way of representing words and Documents with numbers, even though the number sequence is in this case not interpretable.

Finally, the application of Machine Learning techniques, generally available for data analytics or specifically designed for text mining, allows us to discover **sentiment** and **topics** hidden in the text.
1.5. Goals and Organization of this Book

The goal of this book is to give an overview of the whole text mining process and of how to implement it in KNIME Analytics Platform.

We will start of course with importing texts from various sources and in different formats. Chapter 2 is completely dedicated to this topic, including text files, kindle files, social media channels, access to REST APIs, web crawling, and text from forms in web pages. Then in chapter 3 we will cover text-processing techniques: tagging, filtering, stemming, and bag of words extraction. Chapter 4 is dedicated to frequency measures, keyword extraction, and corresponding score calculation.

The first exploratory phase in any data analytics project consists of data visualization, and text analytics is no exception. In chapter 5 the most commonly used text visualization techniques are described. Chapter 6 finally moves to Machine Learning and statistical algorithms for topic detection and classification, while chapter 7 uses Machine Learning algorithms for sentiment analysis.

This book comes with a set of example workflows and exercises. Indeed, when you bought this book you should have received an email with a link to the Download Zone. The Download Zone is just a KNIME file (extension .knar) containing all workflows you need to follow the learning path of this book. Import the .knar file into KNIME Analytics Platform, either via double-click the file or via menu option “File” -> “Import KNIME Workflow” or via right-click LOCAL workspace in KNIME Explorer panel and then “Import KNIME Workflow”.

If the import is successful, you should find in the KNIME Explorer panel a workflow group named TextProcessing_Book with the structure shown in figure 1.6.

The subfolder named “Thedata” contains all data sets used in the following chapters. Each workflow group, named “Chapter …”, contains the example workflows and the exercise workflows for that chapter.

If you are a novice to KNIME Analytics Platform, you will not find much of the basics in this book. If you need to know how to create a workflow or a workflow group or if you still need to know how to create, configure, and execute a node, we advise you to read the first book of this series “KNIME Beginner’s Luck” [4].

There are a few more resources on the KNIME web site about the Text Processing extension.

- Text Processing extension documentation page https://www.knime.com/documentation-3
- Text Processing examples and whitepapers https://www.knime.com/examples
- Text Mining courses regularly scheduled and run by KNIME https://www.knime.com/courses
A number of example workflows can also be found in KNIME EXAMPLES server, at the top of the KNIME Explorer panel, under 08_Other_Analytics_Types / 01_Text_Processing.
From Words to Wisdom

This book extends the catalogue of KNIME Press books with a description of techniques to access, process, and analyze text documents using the KNIME Text Processing extension. The book covers text data access, text pre-processing, stemming and lemmatization, enrichment via tagging, keyword extraction, word vectors to represent text documents, and finally topic detection and sentiment analysis. Some basic knowledge of KNIME Analytics Platform is required. The book has been updated for KNIME Analytics Platform 3.5.

About the Authors

Vincenzo Tursi has been working as Data Scientist at KNIME since May 2016. During this time he worked on text processing, network graph analysis and 360-degree customer data analysis. Before joining KNIME, Vincenzo worked as Business Consultant for Capgemini S.p.A and Business Integration Partners S.p.A. in Italy. He then moved to Germany to work shortly as a Research Associate at Saarland University first and to KNIME later.

Rosaria Silipo has been mining data, big and small, since her master degree in 1992. She kept mining data throughout all her doctoral program, her postdoctoral program, and most of her following job positions. So many years of experience and passion for data analytics, data visualization, data manipulation, reporting, business intelligence, and KNIME tools, naturally led her to become a principal data scientist and an evangelist for data science at KNIME.

ISBN: 978-3-9523926-2-1