DATAQUEST

Amsterdam, Januari 2018

Textmining: from raw text to useful insights

Introduction

Big data is mostly associated with vast amounts of numerical data. Using mathematical models, this data can be transformed to show hidden patterns and new insights. However, most of the data we generate daily is not in the form of numbers, but as written text (just like this whitepaper!). Luckily, computers are nowadays capable of handling texts almost as easy as numbers. A significant part of the tasks we ask computers to perform falls in the category of text mining or the extraction of useful information from written text.

Throughout this whitepaper, we will use the following example to highlight how text mining can be used to add value to your company. Suppose you lead a customer support team whose job it is to answer customer emails. Your team receives thousands of emails per day and, as it turns out, it is impossible to answer all of them by hand. You wonder if computers can help stream line the work flow of your team. Maybe emails can be sorted based on topic and have a dedicated person per topic? Or perhaps emails can be answered by computers altogether. That might sound a bit too dangerous, so maybe you would like to answer emails of angry customers yourself.

Natural language processing (NLP)

Before continuing with our example, let's give some context about text mining. Text mining uses tools from a field called natural language processing (NLP) to derive information from text. Where the concept of text mining is relatively new, NLP finds it origin with the birth of the first computer when people tried to emulate a human response from a computer and tried to link real-world information to computer-understandable data.
The different aspects of NLP can be roughly divided in four categories: 1. Syntax, 2. Semantics, 3. Discourse, and 4. Speech as is shown in Fig. 1.

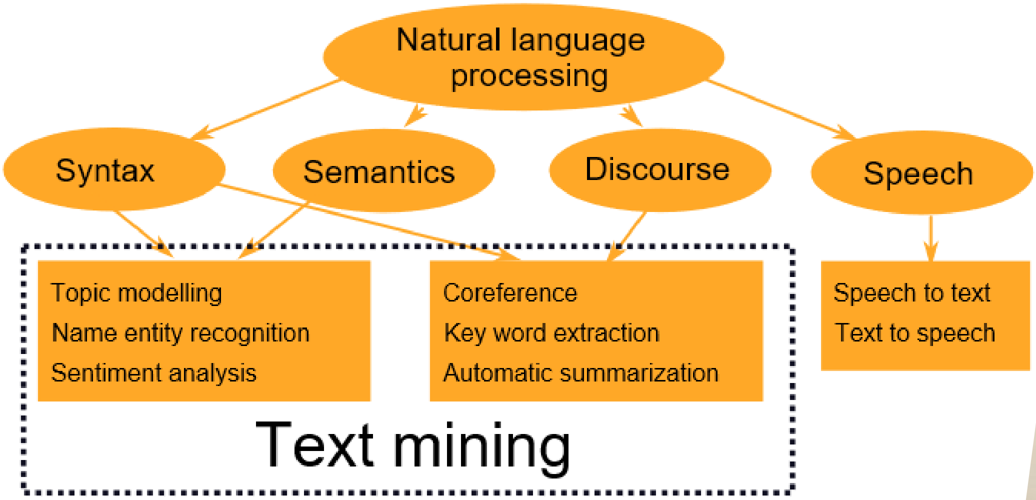


Figure 1. Text mining combines different tools from natural language processing. The most common tasks of text mining are indicated in the dashed box. Using an example, we show the power of these techniques.

Syntax covers most of the pre-processing steps required for further analysis of a text or corpus. Some frequently used steps are: part-of-speech (PoS) tagging, parsing and stemming. Semantics deal with the meaning and context of words and phrases within a sentence, while discourse focuses more on the meaning of paragraphs and the whole document. Finally, speech involves the translation of written text to speech and vice versa. Text mining typically combines tools from these four categories as indicated in Fig. 1. Using our example of the customer support team we show the power of these techniques.

Useful text mining tools

Topic modeling

Let's get back to our example. The idea of separating emails by topic is appealing. Using topic modeling we can determine the main topics of the email and sort them accordingly. Several methods exist to perform this task. When we have stored all previous emails together with the manually assigned topic, we can train a classifier using these old emails and use it for the new, unclassified emails. This is an example of supervised learning (i.e. with labels). When no labelled emails are present, we can use various unsupervised learning techniques to train a model. This does require a significant number of emails.

Name entity recognition and coreference

After successfully classifying the emails, we would like to inspect the content of the messages. Consider the following snippet:

> *I bought a computer, model Fancy2018, for my son last Thursday at the Best Computer Store in Amsterdam. One week later, his computer screen is no longer working.*

With name entity recognition (NER) we can label parts of the sentences as either being a person, a location, a date or time or something else. For example: we may recognize Fancy2018 as a product name, Thursday as a date and Amsterdam as a location. Furthermore, using coreferencing, we can deduce that "his computer screen" refers to the screen of the computer this customer bought for his son. A computer program can perform both these tasks, making it easy to quickly extract key information such as product and store location as well as the relation between phrases.

Key word extraction and automatic summarization

Supplying each email with some key words and a small summary makes searching through the emails a lot faster. We could use NER and coreferencing to provide us with the key words, but there are dedicated algorithms available that use a combination of different techniques, often with great success. Furthermore, a whole summary can be generated by a computer. The summary even comes in two flavors. Extraction-based summaries use key phrases from the document to generate a summary while abstraction-based summaries try to paraphrase the document to capture its essence.

Sentiment analysis

We are now able to classify our emails, provide an overview of important entities mentioned in the emails and have generated key words and a short summary. The final step is to sort out the negative emails and deal with them in person.

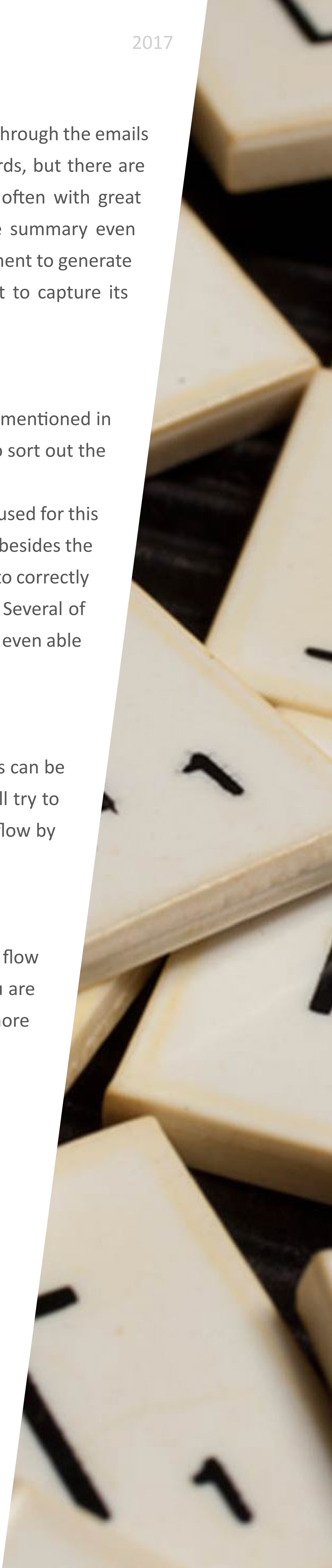
Via sentiment analysis the tone of the email can be extracted. Often the algorithm used for this task is trained on labelled data. A famous example is a set of movie reviews where besides the written text a numerical value is given. These reviews can be used to train a model to correctly identify the sentiment or tone of the review from negative to neutral to positive. Several of these pre-trained models are publicly available ready to be used. Some models are even able to pick out nuances such as the usage of capital letters and exclamation marks.

Next step

With our safety net in place to pick out the negative emails, the categorized emails can be handed over to either your team or perhaps even to a computer program that will try to answer the emails for you! In any case, you have significantly improved the work flow by applying text mining techniques.

Conclusion

We have shown via a simple example how text mining can improve the work flow within a company, but text mining can be applied to many other situations. If you are interested how DataQuest can help you with your text mining needs, you are more than welcome for a cup of coffee at our office.



An example of sentiment analysis

Sentiment analysis on a set of documents works best when the model is trained on similar documents to ensure that the model knows about the jargon and common phrases. This often requires a large collection of documents including labels to indicate if the content is either positive or negative. Fortunately, several pre-trained models exist that can readily be used. One is the VADER sentiment analysis tool (see https://github.com/cjhutto/vaderSentiment). This model is trained using social media such as Twitter and product reviews on Amazon.

When you are a customer support engineer and receive thousands of emails a day, it is impossible to read them all, let alone find the negative emails that require immediate action. Using three examples below, we show how VADER can correctly find the negative email. Consider the following short email:

> *I am very happy with the purchase of my new computer. It works like a charm. My only problem is setting up my email account. Can you help setting up my email account?*

Reading this as a human, we would classify this as a positive email. Indeed, on a scale of -1 to 1 VADER scores this as 0.85 which is rather positive.

The next email

> *Hi, all of the sudden my computer only shows a blue screen when I turned it on. Rebooting the computer did not help and I do not know what to do now. How can I solve this?*

is classified as -0.11, suggesting a neutral email. This corresponds well with human judgment.

Finally, the last email displayed below is clearly a negative email and requires immediate action.

> *Dear Sir/Madam, after several phone calls my laptop is still malfunctioning. I am very angry that a brand-new laptop is not functioning as it should. I am also disappointed that the help-desk was not able to solve this. I would like to talk to management to make a formal complaint.*

Indeed, VADER can correctly identify this email as negative with a score of -0.66. Although we used a model that was trained on a completely different set of documents, it still performs well in terms of finding the negative emails. Implementing such a solution in a corporate setting not only saves time, but allows for a faster response to important emails thereby improving the company's image.

Started as a subsidiary of RiskQuest, a leading Dutch consultancy firm in the financial sector, DataQuest is an Amsterdam based consultancy firm with a broad expertise in modelling any type of data, big or small.

In a data driven society, DataQuest understands how data can be used to gain insight and add value to companies in finance, retail and industry. We can help you with the exploratory phase of determining where value can be added, but we are equally at home with a clearly formulated problem.

You are more than welcome for a cup of coffee at our office to see how DataQuest can help your organization.

DATAQUEST

Herengracht 495, Amsterdam
+31 20 693 29 48
www.data-quest.nl

This report is prepared by DataQuest for general guidance on matters of interest only, and is not intended to provide specific advice on any matter, nor is it intended to be comprehensive. No representation or warranty (express or implied) is given as to the accuracy or completeness of the information contained in this publication, and, to the extent permitted by law, DataQuest does not accept or assume any liability, responsibility or duty of care for any consequences of you or anyone else acting, or refraining to act, in reliance on the information contained in this publication or for any decision based on it. If specific advice is required, or if you wish to receive further information on any matters referred to in this paper, please speak directly with your contact at DataQuest or those listed in this publication. Our general conditions apply to services rendered from us, to our quotations, offers, propositions and calculations.