What is Text Mining?

Text mining (also known as text analysis), is the process of transforming unstructured text into structured data for easy analysis. Text mining uses natural language processing (NLP), allowing machines to understand the human language and process it automatically.

Text mining is an automatic process that uses natural language processing to extract valuable insights from unstructured text. By transforming data into information that machines can understand, text mining automates the process of classifying texts by sentiment, topic, and intent.

In a nutshell, text mining helps companies make the most of their data, which leads to better data-driven business decisions.

Difference between Text Mining, Text Analysis, and Text Analytics?

Text mining identifies relevant information within a text and therefore, provides qualitative results. Text analytics, however, focuses on finding patterns and trends across large sets of data, resulting in more quantitative results. Text analytics is usually used to create graphs, tables and other sorts of visual reports.

Text mining combines notions of statistics, linguistics, and machine learning to create models that learn from training data and can predict results on new information based on their previous experience.

Text analytics, on the other hand, uses results from analyses performed by text mining models, to create graphs and all kinds of data visualizations.

Methods and Techniques

There are different methods and techniques for text mining. In this section, we’ll cover some of the most frequent.

Basic Methods

1. Word frequency

Word frequency can be used to identify the most recurrent terms or concepts in a set of data. Finding out the most mentioned words in unstructured text can be particularly useful when analyzing customer reviews, social media conversations or customer feedback.

For instance, if the words `expensive`, `overpriced` and `overrated` frequently appear on your customer reviews, it may indicate you need to adjust your prices (or your target market!).

2. Collocation

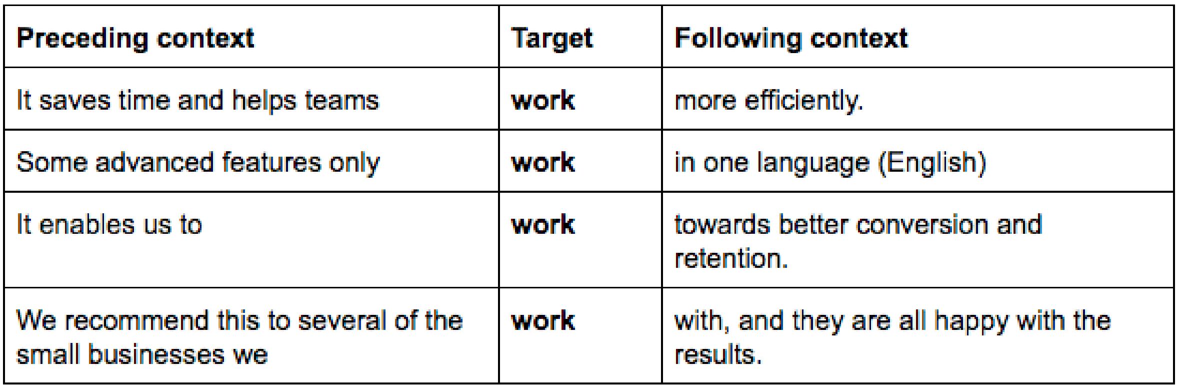
Collocation refers to a sequence of words that commonly appear near each other. The most common types of collocations are bigrams (a pair of words that are likely to go together, like `get started`, `save time` or `decision making`) and trigrams (a combination of three words, like `within walking distance` or `keep in touch`).

Identifying collocations — and counting them as one single word — improves the granularity of the text, allows a better understanding of its semantic structure and, in the end, leads to more accurate text mining results.

3. Concordance

Concordance is used to recognize the particular context or instance in which a word or set of words appears. We all know that the human language can be ambiguous: the same word can be used in many different contexts. Analyzing the concordance of a word can help understand its exact meaning based on context.

For example, here are a few sentences extracted from a set of reviews including the word 'work':



Preceding context	Target	Following context
It saves time and helps teams	**work**	more efficiently.
Some advanced features only	**work**	in one language (English)
It enables us to	**work**	towards better conversion and retention.
We recommend this to several of the small businesses we	**work**	with, and they are all happy with the results.

Advanced Methods

1. Text Classification

[Text classification](#) is the process of assigning categories (tags) to unstructured text data. This essential task of [Natural Language Processing](#) (NLP) makes it easy to organize and structure complex text, turning it into meaningful data.

Thanks to text classification, businesses can analyze all sorts of information, from emails to support tickets, and obtain valuable insights in a fast and cost-effective way.

Below, we'll refer to some of the most popular tasks of text classification – topic analysis, sentiment analysis, language detection, and intent detection.

- **Topic Analysis:** helps you understand the main themes or subjects of a text, and is one of the main ways of organizing text data. For example, a support ticket saying `my online order hasn't arrived`, can be classified as `Shipping Issues`.

- **Sentiment Analysis:** consists of analyzing the emotions that underlie any given text. Suppose you are analyzing a series of reviews about your mobile app. You may find out that the most frequently mentioned topics in those reviews are `UI-UX` or `Ease of Use`, but that's not enough information to arrive to any conclusions. Sentiment analysis helps you understand the opinion and feelings in a text, and classify them as positive, negative or neutral. Sentiment analysis has a lot of useful applications in business, from analyzing social media posts to going through reviews or support tickets. In terms of customer support, for instance, you might be able to quickly identify angry customers and prioritize their problems first.

- **Language Detection:** allows you to classify a text based on its language. One of its most useful applications is automatically routing support tickets to the right geographically located team. Automating this task is quite simple and helps teams save valuable time.

- **Intent Detection:** you could use a text classifier to recognize the intentions or the purpose behind a text automatically. This can be particularly useful when analyzing customer conversations. For example, you could sift through different outbound sales email responses and identify the prospects which are interested in your product from the ones that are not, or the ones who want to unsubscribe.

2. Text Extraction

Text extraction is a text analysis technique that extracts specific pieces of data from a text, like keywords, entity names, addresses, emails, etc. By using text extraction, companies can avoid all the hassle of sorting through their data manually to pull out key information.

Most times, it can be useful to combine text extraction with text classification in the same analysis.

Below, we'll refer to some of the main tasks of text extraction – keyword extraction, named entity recognition and feature extraction.

- **Keyword Extraction:** keywords are the most relevant terms within a text and can be used to summarize its content. Utilizing a keyword extractor allows you to index data to be searched, summarize the content of a text or create tag clouds, among other things.

- **Named Entity Recognition:** allows you to identify and extract the names of companies, organizations or persons from a text.

- **Feature Extraction:** helps identify specific characteristics of a product or service in a set of data. For example, if you are analyzing product descriptions, you could easily extract features like `color`, `brand`, `model`, etc.

Why is Text Mining Important?

Individuals and organizations generate tons of data every day. Stats claim that almost 80% of the existing text data is unstructured, meaning it’s not organized in a predefined way, it’s not searchable, and it’s almost impossible to manage. In other words, it’s just not useful.

Being able to organize, categorize and capture relevant information from raw data is a major concern and challenge for companies. Text mining is crucial to this mission.

In a business context, unstructured text data can include emails, social media posts, chats, support tickets, surveys, etc. Sorting through all these types of information manually often results in failure. Not only because it’s time-consuming and expensive, but also because it’s inaccurate and impossible to scale.

Text mining, however, has proved to be a reliable and cost-effective way to achieve accuracy, scalability and quick response times. Here are some of its main advantages in more detail:

Scalability: with text mining it’s possible to analyze large volumes of data in just seconds. By automating specific tasks, companies can save a lot of time that can be used to focus on other tasks. This results in more productive businesses.

Real-time analysis: thanks to text mining, companies can prioritize urgent matters accordingly including, detecting a potential crisis, and discovering product flaws or negative reviews in real time. Why is this so important? Because it allows companies to take quick action.

Consistent Criteria: when working on repetitive, manual tasks people are more likely to make mistakes. They also find it hard to maintain consistency and analyze data subjectively. Let’s take tagging, for example. For most teams, adding categories to emails or support tickets is a time-consuming task that often leads to errors and inconsistencies. Automating this task not only saves precious time but also allows more accurate results and assures that a uniform criteria is applied to every ticket.

References & Additional Readings:

- https://monkeylearn.com/text-mining/
- https://www.richardtwatson.com/open/Reader/_book/text-mining-natural-language-processing.html
- https://libguides.library.usyd.edu.au/text_data_mining/methods