Text Mining: Theory and Applications

By:
Anurag Nagar

It is estimated that more that 80 percent of all data is in text format, and most of it is unstructured. This text data can be found everywhere - books, news, blogs, emails, stock analysis, social networking, just to name a few. Clearly, there is a need to mine this data and extract useful information from it.

In this tutorial, I will start off by presenting relevant theoretical background of text mining. I will cover topics such as common text mining tasks, special features of text data and how it is different from other forms of data such as numerical data, and how unstructured text is converted to knowledge. I will also cover text processing techniques which allow text data to be converted to document vectors using the Vector Space Model. This will lead to a discussion of document similarity measures and how various data mining algorithms can be used for document classification and clustering.

In the second part, I will show how real-world text mining can be performed using various software tools. Specifically, I will use the **tm** package in **R** and various plugins to show how data can be easily extracted, processed, and summarized according to various specifications. I will also use these packages to extract current news from various sources such as Google News, Yahoo News and Twitter. These sources can also be used to get latest news about various companies whose stocks are traded on major exchanges. By using text mining techniques, we can get a feel of the sentiment prevailing about the company in question. This sentiment has been known to have a strong effect on the stock price. I will use specific terms from the news and try to find how strongly they are correlated with the fluctuation in the stock price.