ABSTRACT

Finding Core Topics in XML Text-based Files using Latent Dirichlet Allocation

Kirushnaamoni Ramakrishnan

Keywords: Text mining, LDA, XML-formated text files

In this presentation, the research methodology used by the researcher with a focus on topic modelling will be described - a method used in text mining to discover common themes or topics in a large collection of text files or documents. This helps to understand what large sets of text files or documents are about without having to read each one individually. As a result, large amounts of text files or documents can be organized into different groups based on the topics they contain.

This is done by an algorithm called Latent Dirichlet Allocation (LDA), which identifies words that can be grouped under a specific topic. Previous research applied LDA methods to text files. In my research, XML-formatted text-based files have been used for it because sometimes the content available on the web is in the form of XML (Extensible Markup Language) files. XML is a method of writing and storing data over the web.

For this study, a dataset consisting of 19,320 XML-formatted text-based blog files was used. The files were divided into six groups, namely: a) Females b) Males c) Age – less than or equal to 20 d) Age – greater than 20 e) Students and f) Everyone. LDA was used to extract 20 topics that comprised ten words each for each group. The goal was to identify the two most common topics among the groups so that innovative products could be manufactured by an innovation company based on the blogger's interests. Apart from that, the sentences that contained these words were extracted from the files to understand if the sentences matched the topics. The presentation concludes by explaining the findings and future work for the study.