THESIS PROPOSAL

Automatic Thema Classification for E-Books

Background

When Storytel receives e-books from publishers, they receive both the book content and metadata. The metadata files come in a format that requires the specification of so-called themas, such as *Classic fiction*, *Russia*, or *Early 19th century*. These themas are defined in a global standard that specifies both general subject classes and more fine-grained thema qualifiers. Today thema codes are manually added to books by the publishers, which is time-consuming and suffers from all sorts of flaws due to different interpretations of the classification system. For some of the books thema information is missing completely.

Project description

The purpose of this thesis project is to explore the use of automatic techniques for assigning thema codes to e-books, given the raw text of the books as the input. The thesis will study different (hierarchical) text classification approaches with a very large number of classes. The student will propose and evaluate different classification methods. As a baseline method an approach similar to (but less complex than) the winning system of the Kaggle hierarchical text classification challenge could be used (Puurula et al., 2014).

Antti Puurula, Jesse Read, and Albert Bifet. Kaggle lshtc4 winning solution. Technical report, 2014.

Customer

Storytel, Stockholm

Contact

Marco Kuhlmann, marco.kuhlmann@liu.se

Student profile

Background knowledge in text mining and natural language processing, via courses such as TDDE09 and TDDE16