# Call for Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

## Automated Classification of Mathematical Documents

The Mathematical Subject Classification Scheme (MSC [1]) is a three-level hierarchical classification scheme for mathemtics and neighbouring areas that is jointly developed and continuously refined by the two large information providers for mathematical publications, zbMATH [2] and MathSciNet [3]. At zbMATH, a product developed by FIZ Karlsruhe‘s Department of Mathematics located in Berlin, the MSC classification is used for structuring the internal workflow as well as for interlinking, filtering, and browsing functionality in the user interface.

Automated classification of text documents in terms of a fixed classification scheme is a classic task in Natural Language Processing. The MSC use case is unique in two respects: One is the use of information and formats specific to mathematics, in particular formulae encoded in LaTeX. The other is the presence of additional structured metadata providing semantic context, namely information about authors, their publication history and coauthor networks, and about journals and their area-specific orientation. Training data is available in the form of about 3.4 million hand-annotated documents.

Your task in this thesis will be to develop a supervised learning algorithm for classification of mathematical text according to the MSC classification scheme using suitable NLP tools able to take the additional information into account. The algorithm should be able to adapt both to additional training data as well as to updates of the underlying classification scheme (the next update driven by communitiy feedback is scheduled for 2020 [4]). The task is expected to require mastery of standard techniques as well as the development of new approaches specific to the use case. In addition to requiring scientific work suitable for a master‘s thesis, the finished algorithm is planned to be put into production at zbMATH after the end of the project.

This thesis will be supervised by **Prof. Dr. Harald Sack, Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe**.

[1] http://msc2010.org/
[2] https://zbmath.org/
[3] https://mathscinet.ams.org/
[3] https://msc2020.org/

zbMATH
the first resource for mathematics

Which prerequisites should you have?

- Good programming skills in Python or Java
- Interest in Machine Learning approaches
- Interest in Deep Learning technologies

Contact person:
**Dr. Fabian Müller**
**fabian.mueller@fiz-karlsruhe.de**