Thesis proposal:
Word Vectors for Swedish
Suitable for Bachelor’s thesis (15 credits) or a project course

Background
How can words be represented in a computer? One way to do this is to map each word into a vector whose dimensions correspond to various syntactic and semantic aspects of the word at hand. Several techniques exist for obtaining word vectors from raw text; one recent example is Google’s word2vec. Word vectors are useful because they support a notion of similarity between words and can be used as the input to neural networks, which facilitate higher-up tasks such as part-of-speech tagging, parsing, sentiment analysis, and machine translation. The goal of this project is to lay the grounds for such research by creating a standard data set with general-purpose word vectors for the Swedish language.

Project description
Your task is to generate and evaluate word vectors trained on large amounts of Swedish text. To generate word vectors you will first need to familiarise yourself with relevant theoretical concepts and software as well as identify and pre-process suitable data sets. To evaluate the trained word vectors you will need to choose appropriate benchmarks and conduct experiments. At the end of the project you will prepare a scientific report, documenting and discussing your findings. Depending on the outcome, the project may be expanded into a Master’s thesis.

Student profile
You should be familiar with the basic principles of language technology (acquiring and pre-processing textual data) and linear algebra (notation for vectors and matrices). You should also be able to program in either Java or Python.

Contact
Marco Kuhlmann, marco.kuhlmann@liu.se