LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Bag-of-words based news recommender system**

Supervisor: Sebastian Gerstner

Examiner: Hinrich Schütze

Level: BSc

Summary: I'm often overwhelmed by the sheer amount of news articles there is to read every day, and as with so many things I find it hard to prioritise. One really handy solution would be a personal recommender system on my own machine: The system would essentially look up scores for the words that appear in the title of a news article, and then give me, say, the top ten articles I should read today. Such a bag-of-words approach is old-fashioned and far from perfect, but highly interpretable: I could just look at my personal word scores and understand how the recommendations come about, learn something about my own preferences, and edit the scores if I feel they are wrong.
Your goal in this thesis will be to implement such a system, including training it on the preferences of at least one user (this can be yourself).
In particular, you will need to make the following design choices:

- Which language?
- What is the unit of analysis? (E.g., orthographic words, subword tokens, morphemes; how to account for stopwords)
- What type of user input should it be trained on? (E.g., click behaviour of the user, preferences from pairs of articles)
- What is the update rule? (E.g., adding $+1$ to the score of every word that appeared in the title of a selected article, or something more sophisticated)
- How should the system appear to the user (during training and during usage)? - Ideally it should have a graphical user interface (GUI).
- Optionally: Is there a way to ensure the system only selects for the topic of the articles and ignores their political leaning? (E.g. treat synonyms with different ideological connotations as equivalent, such as "tax avoidance" vs. "tax optimization")