Statement of Research

Arijus Pleska

October 25, 2016

The project's domain lies in applying machine learning techniques in order to determine structures in metabolomics. Some of the recent research [3] has shown that structures of metabolites can be discovered using topic modelling methodology. Thereby, we will investigate this branch of machine learning to push the current results even further.

We have chosen Latent Dirichlet Allocation (LDA) [2] as our main subject of research. However, we will take into account other techniques such as Probabilistic Latent Semantic Indexing (PLSI) [1] as well. The objective of our research is to optimise the performance of the latter models in order to be compatible with metabolomics. Most importantly, we are going to address the issue of the impact of noise, which is caused by the limitation of the data collection technology. Also, we are going to enhance the currently used mathematical underlying, and suggest appropriate software engineering practises. Further, we will use robust tools to deliver and display results in high quality.

Effectively, the research is expected to increase performance of the discussed models, and to mitigate the previously mentioned obstacles. First of all, we will familiarise with the current research in topic modelling. This preliminary step will allow to familiarise with the terms used in machine learning, and, also, it might suggest alternative approaches in tackling the problem. As a part of the familiarisation process, the models studied will be implemented concurrently with the research. Finally, the results will be documented using adaptive tools. That is, suggested ideas will be written in Jupyter notebook; also, the results will be processed using the Plotly platform to produce interactive plots of high quality.

References

[1] Thomas Hofmann. Probabilistic Latent Semantic Indexing, 1999.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation, 2003.

[3] Justin J.J. van der Hoofta, Joe Wandya, Michael P. Barretta, Karl E.V. Burgessa, and Simon Rogers. Topic Modeling for Untargeted Substructure Exploration in Metabolomics, PNAS 2016.

Refs need journals, pages, volumes etc.

Good first draft but needs some more detail (see below)

General

- fix refs - all need journal name, page number, volume, etc
- You don't mention imaging at all — this is really an imaging problem ie applying topic modelling to Mass Spec Imaging Data
- Need some testable things (i.e. some science!). For example, we are interested in using LDA to incorporate smoothness in the topic contributions at different pixels so we could test the extent to which this improves the reconstruction of topics using data that we generate (i.e. where we know the truth)
  - Make these as specific as possible. i.e. what will you measure and compare between the two approaches (LDA vs Spatial LDA)

- Try to avoid "padding" - i.e. adding text to fill space. e.g first sentence of paragraph 3.