# Semester Project:
# Hierarchical Topic Models for Neuroscience

J.-C. Chappelier
IC/LIA
jean-cedric.chappelier@epfl.ch

R. Richardet
Blue Brain Project (BBP)
renaud.richardet@epfl.ch

29 May 2013

**Abstract**

The main goal of this project is to extract various numerical entities from Scientific Literature. The main focus will be put on protein concentration in specific regions.

## 1 Overview & Goal

Today, a huge amount of very specific data, of interest to various researchers currently engaged in the Blue Brain Project, is hidden in the bodies of scientific publications and cannot be automatically consumed. One goal of the BlueBrain NLP team is to parse scientific articles and extract structured content from them.

The goal of this project is to build and deploy a ready-to-use UIMA component providing hierarchical topics models, building upon the work from a former Master student. One purpose of this UIMA component will be to propose missing synonyms and entries in an ontology, (e.g. NIF's ontology) using state-of-the-art Hierarchical Topic Models (which will be explaines in details). In the final deployment, the component should be able to handle more than 10 millions abstracts of scientific papers or several millions full-text papers.

## 2 Project steps

1. evaluate existing modules
2. integrate existing softwares for hierachical LDA into the BBP-NLP framework
3. evaluate and iteratively improve the models used

# 3 Environment

This semester project is co-supervised between IC-LIA and SV-BlueBrain Project.
Material and working environment will be provided at BBP (QIJ 3).
Existing work:

- Former student's semester project on (non-hierarchical) LDA models
- UIMA modules for preprocessing

# 4 Required Skills

The target student is a typical Computer or Communication Science Master Student, with basic knowledge about Natural Language Processing.

# 5 References


[1] J.-C. Chappelier, Topic-based Generative Models for Text Information Access, In Textual Information Access – Statistical Models, E. Gaussier and F. Yvon eds, ch. 5, pp. 129-178, Wiley-ISTE, April 2012.

[2] Topic Models for Taxonomies, Bakalov 2012 (`http://dl.acm.org/citation.cfm?id=2232861`)

[3] Probabilistic Topic Models for Learning Terminological Ontologies, Wei et al. 2010 (`http://dl.acm.org/citation.cfm?id=2232861`)