Scientific Programming

IT & Cognition

MA

September 2013 – December 2013

Overview

The Scientific Programming course consists of 14 lectures by Anders Søgaard and 14 exercise classes with instructors Peter Halkier and Claire Joyce. The exam is 4-6 assignments you submit throughout the semester.

Purpose

The purpose of the Scientific Programming course is to provide students with little or no or considerable programming experience with the tools to implement scientific experiments relating to cognitive technologies, as well as with methods for quantitative data analysis and visualization.

Lectures

At the end of each Friday lecture the students are presented with exercises for the following Thursday. These exercises can be solved individually or in reading groups, and with the assistance of their instructors on Tuesdays. A select subset of these assignments will constitute the exam. PP is the Python Programming book; CV is the Computer Vision book. PP: http://upload.wikimedia.org/wikipedia/commons/9/91/Python_Programming.pdf CV: http://programmingcomputervision.com/downloads/ProgrammingComputerVision_CCdraft.pdf The lectures are:

1. Toolbox: Functions. Topics: Data conversion, representing text, from folders to BOW. Modules: glob. Reading: PP9, PP10, PP13, and PP14. Exercise: Install NLTK and text2bigrams.py.
2. Toolbox: Classes. Topic: Handwritten digit recognition and BOWs in scikits. Modules: scikits. Reading: PP18 and PP19. Check this out: http://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html#example-neighbors-plot-classification-py and https://thenewcircle.com/s/post/1152/scikit-learn_machine_learning_in_python Exercise: Nearest neighbor using lists.
3. Toolbox: arrays, matrices. Topic: Nearest neighbor using scipy. Modules: numpy, scipy. Reading: http://www.engr.ucsb.edu/~shell/che210d/numpy.pdf Exercise: Implement and evaluate Rocchio on both texts and images, including error analysis (**exam**).
4. Topic: Implementing (averaged) perceptron. Reading: http://scikit-learn.org/stable/modules/linear_model.html Exercise: Implement stochastic gradient descent.
5. Topic: Evaluation against gold data. Modules: pylab. Reading: http://matplotlib.org/users/pyplot_tutorial.html https://thenewcircle.com/s/post/1133/advanced_matplotlib_tutorial_with_library_author_john_hunter Exercise: Learning curves on your own document classification dataset (**exam**).
6. Toolbox: scipy.stats, pandas. Topic: Naïve Bayes, distributional assumptions. Reading: http://www.youtube.com/watch?v=DXPwSiRTxYY http://pandas.pydata.org/ Exercise: Statistical analysis of features in 20 newsgroups (oral presentation). Implement feature selection.
7. Toolbox: PIL. Topic: Image Processing. Reading: CV1, CV2, CV9. Exercise: Rank images by brightness (**exam**).
8. Toolbox: scikits, pyfaces. Topic: Face recognition and clustering. Reading: CV6, http://pyfaces.blogspot.dk/ Exercise: Reading groups present a face recognition experiment in class. Implementation and dataset of your own choice.
9. Toolbox: scikits, theano. Topic: Dimensionality reduction. Reading: http://scikit-learn.org/stable/modules/decomposition.html Exercise: Train auto-encoder (neural network) to learn low-dimensional representations.
10. Topic: Sequential labeling. Reading: http://lxmls.it.pt/2013/guide.pdf (Ch. 2). Exercise: Implement a simple HMM and compare your performance on PTB with an off-the-shelf

implementation (**exam**).

11. Topic: Interactive Python. Reading: TBA. Exercise: Design an annotation script for sequential labeling of words.
12. Toolbox: nltk, twittersearch, etc. Topic: Collecting web data. Exercise: Collect a corpus of tweets and annotate them with POS (**exam**).
13. Guest lecture: My favorite Python snippets. Speaker: TBA. Reading: TBA. Exercise: Design a scientific experiment with your group, including statistical significance testing – to be presented by another reading group.
14. Topic: Python and scalability. Reading: https://thenewcircle.com/s/post/1094/python_in_big_data Exercise: Student presentations.

Evaluation

The evaluation is pass/fail based on active participation (the 4-6 exam assignments).