***DL - Code***

**Scope:**

Module: NumPy

Technologies: Python (scrapy, spark, Keras/Tensorflow)

Requirements:

- program learns from structured, clean datasets.

- program takes questions in natural language (English) and responds with code

- program should not provide not runnable code

**Milestones:**

1. Extract and save data
   1. From:

- Stack overflow

- Python documentation

- NumPy Documentation

- Gist, GitHub, etc.

- pretty Code examples (many comments)

* 1. How to extract

- Stackoverflow dump

- Python webcrawler (e.g. scrapy)

- (Stackoverflow) APIs

* 1. Find Data Model

save as:

- JSON ?

- MongoDB ?

- SQLite ?

1. Clean Data

- filter unnecessary words (e.g. spark wordcount, python script) OR

- NN for classification of words (relevant/irrelevant)

- Code checks for validity: determine how to do checks

- unify variable names

- seperate example "setup" from the core code functionality

1. Attempt to train Recurrent Neural Network for

(to be determined)

- API sequences (like in the paper Deep API learning)

- Complete code fragments

- Extracting variable types, or even roles

1. Use the NN

- command line interpreter

- Webinterface or PlugIn

**Feedback:**

* Feeding an RNN with data should be simple
* Unification
* Prepare data
* Clean code