Goal of the project

1. Main idea: mining programing language data, learn the functionality of a code segment.
2. Entry level goal: classify the code functionality from a set of finite abstract descriptions, e.g. one word.
3. Advanced level: Use detailed descriptive sentences to address the functionality of given code segment.

The set of source selected for the project

For this problem, we will focus on codes written in Python. We selected two data sources.

1. Well developed and widely used Python packages.
   a. They are well-developed and follow certain standard of software engineering.
   b. Functions are usually well-commented and/or well-documented.
   c. Comments and documentations represent the usage and purpose.
2. Codes from website stack overflow.
   a. Question title and description is a text document which gives detailed description of the codes in following answers.
   b. Codes provided in the different answers for the same question will fall into same category.

Method used to extract structured data from the data sources.

1. For data from installed Python packages.
   a. We can use Python built-in and/or third-party program analysis libraries to extract:
      i. Package name,
      ii. Version information,
      iii. Function usage description (e.g. docstring) as text document,
      iv. Input parameter number,
      v. Return type,
vi. Other structural data that may be useful.

b. Learn how to and enjoy the difficulties playing with source code as data.

2. For data from Stack Overflow website. Use “scrapy” package to crawl question pages.
   a. Search question title from pages under Python label.
   b. For each question title, check the answers.
      i. Get question description
      ii. Get answer code, votes, answer comments
      iii. Other structural data that may be useful.
   c. Learn the tweaks and avoid being banned. :)

● Information extracted from the text documents

1. `docstring` of functions from installed Python package.
   a. Function’s functionality
   b. Version information
   c. Related functions.

2. Question description on Stack Overflow.
   a. Intention and functionality of codes answered below.
   b. Specific difference/discrimination between similar functions

3. Answers and comments on Stack Overflow.
   a. Function’s functionality
   b. Usage

● Open-source tools

- scrapy: web crawling
- NLTK: natural language processing
- numpy, scipy: scientific computation
- scikit-learn: machine learning
- Keras, Tensorflow/Theano: deep learning
- Pylint: static analysis with Python