**Programming with Python with Data Science**

**Train the Trainer Program End Project**

**Project Title**: Develop a SEO tool to analyse live web pages

Search Engine Optimization (SEO) is an important aspect of a web page to gain importance for a search engine to be able to display it earlier in the search list. The optimization is based on a lot of factors such as title, description, header tags and keyword density. Different search engines will have their own mechanisms that calculate the score of a keyword on the page and thus work out its ranking in the search order.

The course project is to develop a generic SEO toolset which will help a web developer analyse live web pages for keywords and other components of the page which contribute to SEO. We should be able to provide the keywords of interest and understand its density pattern across the various components of the HTML page. The web page analysis should be developed to be done in a batch mode where we can analyse hundreds of web pages and the results should be saved as reports in spread sheets with graph presentations where necessary.

The project would not only give exposure to a real life problem solving using Python, but would also make learners understand about Internet and Search Engines and how they relate to each other.

Requires Python 2.7+, BeautifulSoup4, minidom, nltk, numpy and urllib2.

### Installation

#### PIP

pip install -r requirements.txt

#### Python Shell

>> import nltk

>> nltk.download()

#### Usage

If you run without a sitemap it will start crawling at the homepage

./analyze.py http://www.domain.com/

Or you can specify the path to a sitmap to seed the urls to scan list.

./analyze.py http://www.domain.com/ path/to/sitemap.xml

## Requirements

* BeautifulSoup4
* nltk
* numpy
* requests