International Institute of
Information Technology, Bangalore

Project Proposal
CS/DS 706 Machine Learning

Python Stackoverflow QA Analysis

Akanksha Dwivedi - MT2016006
Tarini Chandrashekhar - MT2016144

Instructor :
Prof. Dinesh Babu J

August 30, 2017

Contents

1 Brief Description

This project aims at utilising natural language processing and exploratory analytics on a dataset consisting of Python question and answers on Stack Overflow, to design a novel course structure for teaching Python.

1.1 Problem Formulation

2 Dataset

The dataset consists of full text of questions and answers from StackOverflow, that are tagged with the python tag, useful for natural language processing and community analysis.

This is organized as three tables i.e three .csv files:

- **Questions** contains the title, body, creation date, score, and owner ID for each Python question.
- **Answers** contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
- **Tags** contains the tags on each question besides the Python tag.

3 Proposed Plan of execution

4 Main Challenges

5 Learning Objectives