SMAI Project
Progress Report

Title: Competitve problem tag generation and similarity analysis

Team Members:

Name: Chittaranjan Rath
Roll No: 2018201007

Name: Nitish Srivastava
Roll No: 2018201012

Name: *Prakash* Nath Jha
Roll No: 2018201013

Name: *Suraj Garg*
Roll No: 2018202003

Objective:

To predict tags(MultiLabel Prediction) for the provided input set(combination of problem statement and solution for cp questions).

Considered Approaches:

- Building a model considering multiple(all) tags
- Building separate model for each individual tags

Considered Models:

- Building basic clustering models
  - linear svc
  - Mutinomial Naive Bayes
  - Logistic Regression

Deliverables:

- Scraping of various websites like codechef,codeforces for data collection.
- Preprocessing of data for model usage
- Analysis based on above approaches and models

Technologies/libraries to be used:

- Tensorflow
- python 3
- scikit learn
- nltk
- BeautifulSoup

Github Repository :

https://github.com/gargsuraj12/SMAI_FinalProject

Probable DataSets:

- *https://www.dropbox.com/sh/k4kggx0m9d9r7qn/AAArYSiGkMsr_n8zQNejc04Va?dl=0*

Points to be discussed:

- Till date we have fetched approx 5000 sample inputs(cp problems with solution and tags,etc.) Do we need to consider more data extraction?
- Recommended models for these kind of classification
- How NLP methods can be applied on solution extracted
- Will addding new features to our dataset help?