Implementation of ML Framework to Prioritize Meaningful Changes In Environment Related Websites

Student

Name : Ankur Shukla
Affiliation : Pursuing Geoinformatics and Natural Resources Engineering, at Indian Institute of Technology, Bombay as Masters Degree
Location : Mumbai, India, UTC +5:30
GitHub : daas-ankur-shukla
Slack handle : daas-ankur
Email : work.ankurshukla@gmail.com
Other Contact:https://www.linkedin.com/in/ankur-shukla-8b73b266/

Project Details

Project Title:
Implementation of ML Framework to prioritize meaningful changes in Environment related websites

Project Abstract/Summary:

The project aims to develop and implement a machine learning framework to prioritize diffs and generate appropriate actions.

Describe the need your project fills:
This project aims to fulfill the following objectives to prioritize the webpage diffs:

1. Conversion of html diffs to a common schema for machine learning algorithms
2. Machine Learning model to accept formatted diffs as input and output their priority
3. Integrate the model and other components with other EDGI components to take human analyst feedback for improving the model.

Describe how your project meets these needs:
A website can undergo changes at various levels, viz: its design, layout, content and interface. For this project we are concerned with changes in the content of a website. Further the content of a website can be categorized as images, text (headings, paragraphs). In this project I propose to deal with text changes in a website.

To get the most out of machine learning models, we first have a definite diff format which can be fed to ML models. Thus the first focus of my project will be to arrive at a schema for converting various diff formats to a common format.

1. Conversion of html diffs to a common schema to form the feature matrix for machine learning model

   After thorough discussion, we can arrive at a feature matrix, with the best possible feature specification to train our ML model. The following can be a prospective list of features (non-exhaustive):

   a. HTML Tag difference
   b. Word Difference
   c. Position Difference

   There shall be intermediate vectors for each class of change. For example for HTML tags the following may be a priority vector (non-exhaustive):

Tag	h1	h2	b	p
Priority	4	3	2	1

   These priority vectors can help us give a relative severity index to a change derived from a diff. For example, a change from <p> tag to <h1> is a more severe change as compared to a change from <b> to <h2>. There will be similar vectors for severity for different change types. These change severity vectors can then be given to a machine learning classifier to prioritize them

2. The severity vectors can then be provided to a multi-class classifier such as multinomial logistic regression or neural networks, to assign priorities to these vectors.
3. The trained models then need to output the prioritized result to other EDGI services so that high priority changes can be analyzed by the experts.
4. The model can also be deserialized, to retrain it or to update their weights to account for some wrong result. This can be done by integrating a user feedback mechanism which can override the priorities assigned by the model. Any changes to the priority vector made by an analyst can be used to retrain the model and improve its weights.

These tasks will be conducted by a framework under which all the components would work. The flow of control for the framework will be:

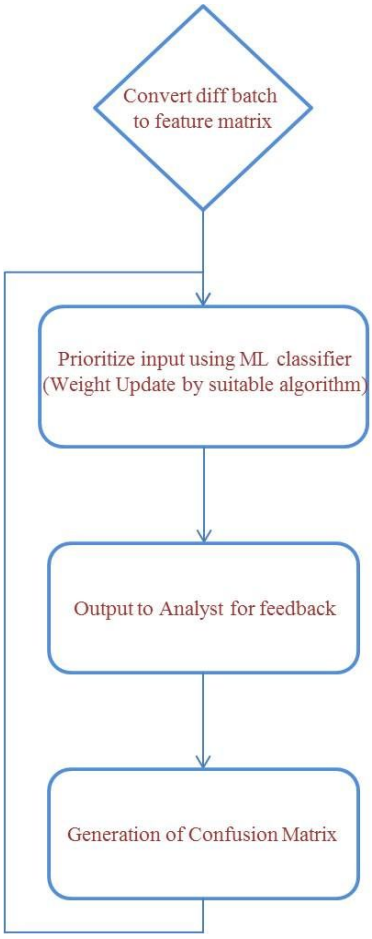

Figure 1: Flowchart showing Framework Functions

Lastly the project will be motivated towards integrating the proposed framework into EDGI services.

Milestones/Timeline:

A. Community Bonding Period (May 4-May 29, 2017):
   a. Make a wiki page on the EDGI repository for fortnightly tasks and report list
   b. Understand the diff format from various website monitoring services (Pagefreezer, Versionista)
   c. Discuss the feasibility of different machine learning algorithms for prioritization tasks.
   d. Discuss the outline of the framework and its functionalities in detail

e. Discuss schema for conversion of diff data to feature matrix for classification algorithm

B. Coding Period (May 30-August 29, 2017):
   a. Coding Period Phase 1 (May 30-June 29, 2017):
      i. Week 1 (May 30-June 5):
         1. Analyze diff data generated from different API
         2. Finalize feature matrix design
      ii. Week 2 & 3 (June 6-June 19):
         1. Make code to map diff data into feature matrix
      iii. Week 4 (June 20-June 29):
         1. Test, debug and document code
         2. Prepare work for Phase 1 submission along with a brief Phase 1 report
         3. Prepare Pre-Phase 2 synopsis
   b. Coding Period Phase 2 (June 30-July 27, 2017):
      i. Week 5 & 6 (June 30-July 13):
         1. Discuss the merits and demerits of different machine learning libraries and zero in on one.
         2. Code a classifier for prioritization task (may be a logistic regression or a neural network)
      ii. Week 8 (July 21-July 28):
         1. Begin implementing a framework on top of the classifier for the following features (non-exhaustive):
            a. Store and retrieve models
         2. Prepare elaborate documentation and tests for Phase 2 components
         3. Prepare work for Phase 2 submission
         4. Prepare Pre-Phase 3 synopsis
   c. Coding Period Phase 3 (July 29-August 29, 2017):
      i. Week 9 & 10 (July 29-August 4):
         1. Finish implementing the framework with the following features (non-exhaustive):
            a. Retrain model by retrieving it and training it with new batch analyst feedback
            b. Query test results for batch of diff data
            c. Output the query result in specified format
         2. Connect the framework to other EDGI services
      ii. Week 11 & 12 (August 5-August 25):
         1. Test, debug code
         2. Write pending documentation

3. Prepare Final Phase submission along with a detailed final phase report

Deliverables:

At the end of GSOC period I propose to deliver the following components:

1. Code to convert diffs to a common format
2. Code for Machine Learning classifier
3. Framework code on top of the classifier to perform above mentioned functions
4. Proper documentation of each component

Resources:

During my project, the following resources will play a crucial role and will facilitate my project work:

1. People:
    a. Environmental Data Experts and Analyst, for their opinion to zero in on the best possible criteria for meaningful changes in a website
    b. Webpage designers for their opinion on which kind of impact a change in a web page can have on a user
    c. People having experience with web monitoring services to help in quick understanding of diff generating APIs
2. Documentation:
    a. Documentation of web monitoring services being used
    b. Documentation of other EDGI components for coding and integration process
3. Literature:
    a. Machine learning literature for designing the classifier
4. Sample Data:
    a. Training data of web page changes (diffs) for training the model

Setup:

- To develop familiarity with EDGI codebase I have setup web-monitoring-db locally.
- For the GSOC period, I will setup all the components required to understand the codebase of already existing components and to implement my proposal.

Ongoing involvement:

- I would say that I have had a fair amount of involvement in the project up till now.
- I have been participating in discussions on slack channel (gsoc & dev) to provide input from my side and receive insights regarding the project. I have also tried suggesting on github issue #18.

- I have opened an issue (#17) concerning the installation instruction provided on the repo. To resolve the same issue I have opened a PR (#18) and I am working on it.

Student Experience

Previous Experience:

- **Work Experience (2014-2015):** Assistant System Engineer at Tata Consultancy Services Ltd
  - Trained in Core and Advanced Java
  - Developed a tool in Java to check for anomalous data in a Hive database
- **Open Source Contribution:** My Github Repo - daas-ankur-shukla
  - Developed a MATLAB GUI to implement low pass filters on a image and extract low frequency component of an image to achieve smoothing effect.
  - Helped convert a C++ code to STL for 2.4 release of PgRouting (PR)
- **Coding Skills:**
  - Python:
    - Image Processing Libraries such as GDAL
    - Django, Flask
  - MATLAB:
    - Simulink
    - GUIDE
  - Java:
    - Core and Advanced concepts
    - Struts, Hibernate
  - Javascript:
    - OpenLayers 3.0

Open Source Project(s) you are working on or would like to:

- Developing a web app using Django, as a smart parking solution based on IoT and PgRouting. I have presented an abstract of the project in FOSS4G ASIA 2017 conference at IIIT Hyderabad. The same work as been chosen for an international publication by the conference community.
- Developing a web application (repo) based on Flask, to allow students to conduct satellite image processing experiments online.
- Apart from these I would love to contribute to projects such as PyOpenCL, Tensorflow and CLTK

Teamwork:

- I am a member of a team of 4 people working on coding a convolutional neural network for classification of human facial expression based on a image dataset.

- I am a member of a team of 3 people in development of the smart parking solution mentioned above.

Interests:
I am an open source & python enthusiast and keen on learning machine learning concepts. Apart from language processing, I am interested in applying machine learning concepts to enhance satellite image processing and interpretation methods. I am also interested in GIS applications. Keen on implementing python based GIS applications.

Commitment:
During the GSOC period I will be able to work for 7 hours a day from May till June end. From July, my college will resume. Then onwards I will be able to devote 4 to 5 hours on weekdays and about 7 hours on weekends, making up to about 40 hours a week.