International Institute of
Information Technology, Bangalore

Project Proposal
CS/DS 706 Machine Learning

Python Stack Overflow Q&A Analysis

Akanksha Dwivedi - MT2016006
Tarini Chandrashekhar - MT2016144

Instructor :
Prof. G. Srinivasaraghavan

August 31, 2017

Contents

1 Brief Description

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

1.1 Problem Formulation

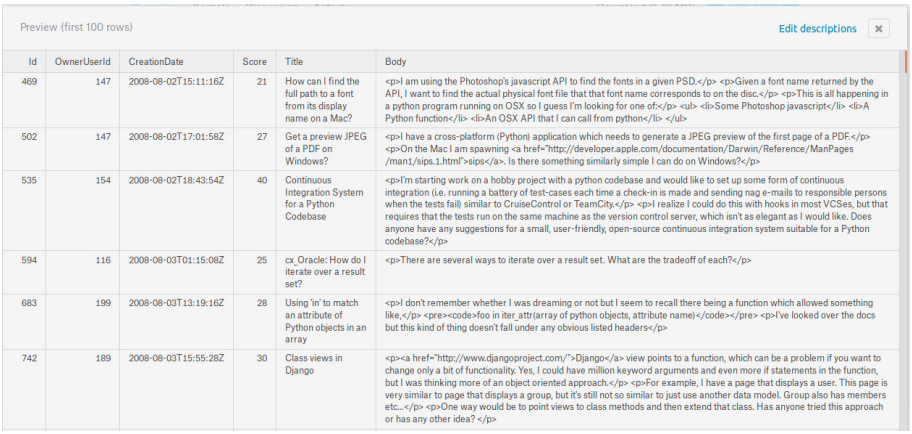
- Application/System (Python Course Design)
  - MP Module (Exploratory Analytics and Natural Language Processing on Questions and Answers)
    - ML task (Unsupervised Text mining)
      - Features, Models, Optimization algorithm (to be decided based on empirical observations)

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. The dataset is collected over a period of 8 years, from 2008-2016.

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

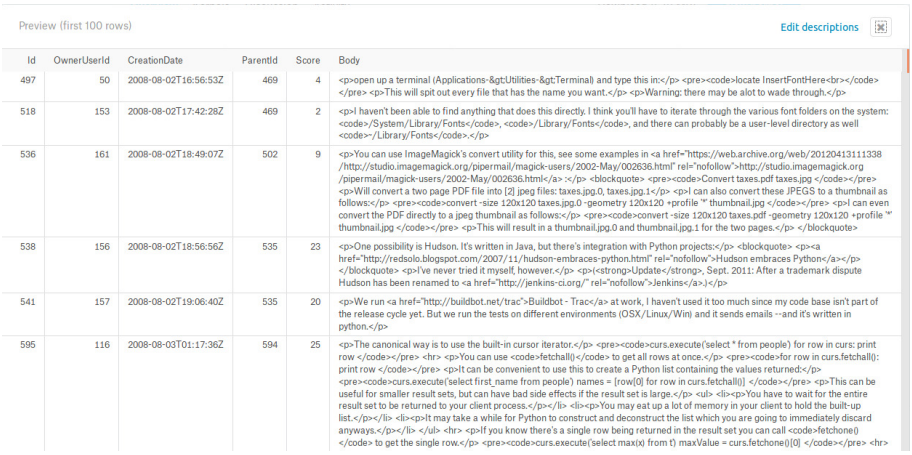
- **Questions** contains the title, body, creation date, score, and owner ID for each Python question.
- **Answers** contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
- **Tags** contains the tags on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



Preview (first 100 rows) — Edit descriptions

Id	OwnerUserId	CreationDate	Score	Title	Body
469	147	2008-08-02T15:11:16Z	21	How can I find the full path to a font from its display name on a Mac?	I am using the Photoshop's javascript API to find the fonts in a given PSD. Given a font name returned by the API, I want to find the actual physical font file that that font name corresponds to on the disc. This is all happening in a python program running on OSX so I guess I'm looking for one of: - Some Photoshop javascript - A Python function - An OSX API that I can call from python
502	147	2008-08-02T17:01:58Z	27	Get a preview JPEG of a PDF on Windows?	I have a cross-platform (Python) application which needs to generate a JPEG preview of the first page of a PDF. On the Mac I am spawning sips. Is there something similarly simple I can do on Windows?
535	154	2008-08-02T18:43:54Z	40	Continuous Integration System for a Python Codebase	I'm starting work on a hobby project with a python codebase and would like to set up some form of continuous integration (i.e. running a battery of test-cases each time a check-in is made and sending nag e-mails to responsible persons when the tests fail) similar to CruiseControl or TeamCity. I realize I could do this with hooks in most VCSes, but that requires that the tests run on the same machine as the version control server, which isn't as elegant as I would like. Does anyone have any suggestions for a small, user-friendly, open-source continuous integration system suitable for a Python codebase?
594	116	2008-08-03T01:15:08Z	25	cx_Oracle: How do I iterate over a result set?	There are several ways to iterate over a result set. What are the tradeoff of each?
683	199	2008-08-03T13:19:16Z	28	Using 'in' to match an attribute of Python objects in an array	I don't remember whether I was dreaming or not but I seem to recall there being a function which allowed something like, <pre><code>foo in iter_attr(array of python objects, attribute name)</code></pre> I've looked over the docs but this kind of thing doesn't fall under any obvious listed headers
742	189	2008-08-03T15:55:28Z	30	Class views in Django	Django view points to a function, which can be a problem if you want to change only a bit of functionality. Yes, I could have million keyword arguments and even more if statements in the function, but I was thinking more of an object oriented approach. For example, I have a page that displays a user. This page is very similar to page that displays a group, but it's still not so similar to just use another data model. Group also has members etc... One way would be to point views to class methods and then extend that class. Has anyone tried this approach or has any other idea?

Figure 1: Questions.csv



Preview (first 100 rows) — Edit descriptions

Id	OwnerUserId	CreationDate	ParentId	Score	Body

497	50	2008-08-02T16:56:53Z	469	4	open up a terminal (Applications->Utilities->Terminal) and type this in: <pre><code>locate InsertFontHere
					</code></pre> This will spit out every file that has the name you want. Warning: there may be alot to wade through.

518	153	2008-08-02T17:42:28Z	469	2	I haven't been able to find anything that does this directly. I think you'll have to iterate through the various font folders on the system: <code>/System/Library/Fonts</code>, <code>/Library/Fonts</code>, and there can probably be a user-level directory as well <code>~/Library/Fonts</code>.
536	161	2008-08-02T18:49:07Z	502	9	You can use ImageMagick's convert utility for this, see some examples in http://studio.imagemagick.org/pipermail/magick-users/2002-May/002636.html : <blockquote> <pre><code>Convert taxes.pdf taxes.jpg </code></pre> Will convert a two page PDF file into [2] jpeg files: taxes.jpg.0, taxes.jpg.1 I can also convert these JPEGS to a thumbnail as follows: <pre><code>convert -size 120x120 taxes.jpg.0 -geometry 120x120 +profile '*' thumbnail.jpg </code></pre> I can even convert the PDF directly to a jpeg thumbnail as follows: <pre><code>convert -size 120x120 taxes.pdf -geometry 120x120 +profile '*' thumbnail.jpg </code></pre> This will result in a thumbnail.jpg.0 and thumbnail.jpg.1 for the two pages. </blockquote>
538	156	2008-08-02T18:56:56Z	535	23	One possibility is Hudson. It's written in Java, but there's integration with Python projects: <blockquote> Hudson embraces Python </blockquote> I've never tried it myself, however. (Update, Sept. 2011: After a trademark dispute Hudson has been renamed to Jenkins)
541	157	2008-08-02T19:06:40Z	535	20	We run Buildbot - Trac at work, I haven't used it too much since my code base isn't part of the release cycle yet. But we run the tests on different environments (OSX/Linux/Win) and it sends emails --and it's written in python.
595	116	2008-08-03T01:17:36Z	594	25	The canonical way is to use the built-in cursor iterator. <pre><code>curs.execute('select * from people') for row in curs: print row </code></pre> <hr> You can use <code>fetchall()</code> to get all rows at once. <pre><code>for row in curs.fetchall(): print row </code></pre> It can be convenient to use this to create a Python list containing the values returned: <pre><code>curs.execute('select first_name from people') names = [row[0] for row in curs.fetchall()] </code></pre> This can be useful for smaller result sets, but can have bad side effects if the result set is large. - You have to wait for the entire result set to be returned to your client process. - You may eat up a lot of memory in your client to hold the built-up list. - It may take a while for Python to construct and deconstruct the list which you are going to immediately discard anyways. <hr> If you know there's a single row being returned in the result set you can call <code>fetchone()</code> to get the single row. <pre><code>curs.execute('select max(x) from t') maxValue = curs.fetchone()[0] </code></pre> <hr>

Figure 2: Answers.csv

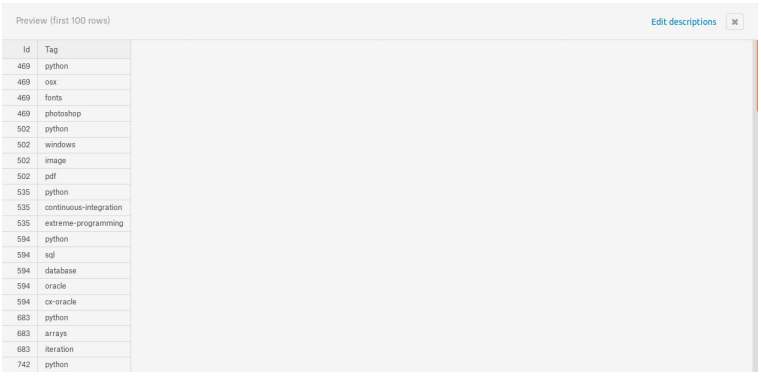


Figure 3: Tags.csv

3 Proposed Plan of execution

3.1 Milestone 1

- Topic association on a rudimentary level by clustering techniques.
- Semantic understanding of the dataset.

3.2 Milestone 2

- Performing various techniques of Unsupervised Text Analysis on the dataset.
- Course structure design based on the filtered dataset, where filtered dataset would be a result of multiple exploratory questions, e.g. most asked questions, most discussed questions etc.

3.3 Milestone 3

- Augment the course structure designed further to a virtual classroom, with attributes like highest rated users as mentors, most discussed questions as interactive sessions, and so on.

4 Main Challenges

- The dataset collected is over a long time period, i.e. 8 years, hence relevant data filtering would be a challenge.
- A very small percentage of questions on Stack Overflow are closed, so we do not have closing timestamp of the questions. We cannot explore the lifetime of a question without closing time.
- Judging the merit of a low scoring answer, as genuinely being irrelevant, being structurally weak or any other inexplicable reason.

5 Learning Objectives

- Experimenting with various unsupervised text mining techniques.
- Semantic understanding of the text through various Natural Language Processing techniques.
- Understanding programming language trends on an active Q&A platform such as Stack Overflow.