Understanding institution level performance (compared to VTU level) using data analytics techniques

Internship report

by

Bhaskarjyoti Das
CSE M tech, Sem 3, 2015
Rajiv Gandhi Institute Of Technology, Bangalore
1RG14SCS03, 10th September,2015

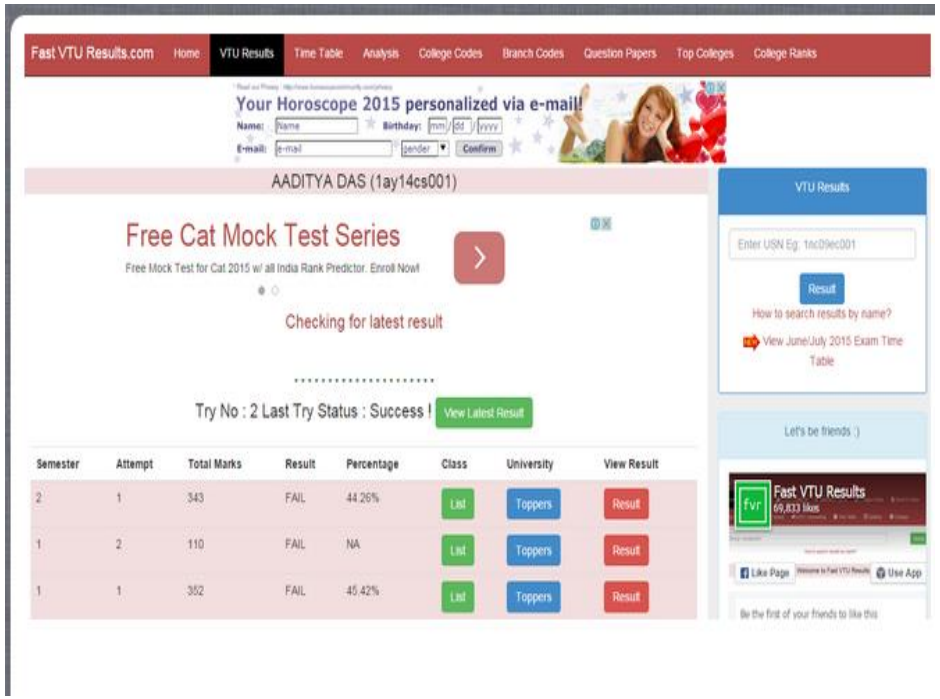
Content

- Goal/motivation
- System design and implementation
- Learning outcome
- Sample output based on 2015 August results
  - Result analysis of 4th Semester (CSE B Tech)
  - Result analysis of 6th Semester (CSE B tech)
  - Result analysis of 2nd Semester (CSE M Tech)

Goal/motivation

- Do a practical data analytics project
- Currently, we cannot compare performance of an institution (RGIT) in a specific subject with VTU level performance
- VTU does not publish university level metrics
  - Get the VTU level data by web scraping
  - Get the subject specific performance data from scraped web pages
  - Do standard statistical analysis (mean, maximum, minimum, standard deviation) for each subject across VTU
  - Compare the above with institution (example –RGIT) level performance

Key implementation issues -1
Choose source of data



http://www.fastvturesults.com/

We get the #no of students and manually find largest USN in a college from this website

http://results.vtu.ac.in/

We get the actual result here..

Key implementation issues -2

- **Choose scraping frameworks**
  - A framework that allows to programmatically fill in web form and get the HTML web page
    - Python Mechanize framework (it simulates a browser)
  - A framework to scrape the HTML web page by making it into a parse tree and search the parse tree (for data) by HTML tag
    - Python Beautiful Soup framework

Key implementation issues -3

- **How to scrape for approx. 10000 USN numbers for each batch?**
  - Programmatically generate the USN numbers as per USN number format for scraping program
  - USN = College code + batch + stream + USN digits (2 or 3)
  - Input -> College code , maximum USN Number

Key implementation issues -4

- **Scraping issues**
  - There is no authentic listing of candidates' USN by VTU
  - Scraping is a semi-automatic error prone process
  - There are some result sheets that are not really applicable or incomplete
  - Scraping fails for many reasons beyond our control (remote server dropping connection, ill-formatted HTML etc.)
  - Scraping is time–consuming : one B tech batch took around 2-3 days of constant effort

- **Dataset issue**
  - We may not 100 % of result data set but will get high 90%
  - For CS 4th semester, we got **10893 complete unique result sheet** though we expected around 12000 based on fastvturesults
  - This mismatch of 1000 is based on many blank result, incomplete results, some defective results and some results we might have left out while scraping

- **Will it affect the statistics much?**
  - No ! as we are calculating mean over 10000+ data points

Some examples of data challenges

August 29, 2015

RESULTS (PROVISIONAL)

MUHAMMED IRSHAD (1AM13CS404)

Semester: 4 Result: FAIL

Subject	External	Internal	Total	Result
Engineering Mathematics - IV (10MAT41)	0	15	15	A
Design and Analysis of Algorithms (10CS43)	48	10	58	P
Unix and Shell Programming (10CS44)	36	16	52	P

Semester: 3 Result: FAIL

Subject	External	Internal	Total	Result
Engineering Mathematics-III (10MAT31)	37	15	52	P
Logic Design (10CS33)	45	19	64	P
Discrete Mathematical Structures (10CS34)	0	16	16	A
Object Oriented Programming with C++ (10CS36)	42	8	50	P

Tota

August 29, 2015

RESULTS (PROVISIONAL)

SYED RIZWAN (1BO13CS403)

Semester: 6 Result: FAIL

Subject	External	Internal	Total	Result
Management & Entrepreneurship (10AL61)	35	22	57	P
Unix System Programming (10CS62)	35	17	52	P
Compiler Design (10CS63)	28	22	50	F
Computer Networks - II (10CS64)	44	18	62	P
Computer Graphics & Visualization (10CS65)	46	15	61	P
Programming Languages (10CS666)	44	16	60	P
Computer Graphics & Visualization Lab. (10CSL67)	36	16	52	P
Unix System Programming & Compiler Dsgn.Lab. (10CSL68)	29	23	52	P

Semester: 5 Result: SECOND CLASS

HETHAN C (1AT13CS401)

Semester: 4 Result: FAIL

Subject	External	Internal	Total	Result
Advanced Mathematics - II (MATDIP401)	0	0	0	A
Engineering Mathematics - IV (10MAT41)	20	15	35	F
Graph Theory and Combinatorics (10CS42)	35	15	50	P
Design and Analysis of Algorithms (10CS43)	28	19	47	F
Unix and Shell Programming (10CS44)	42	20	62	P
Microprocessors (10CS45)	49	18	67	P
Computer Organization (10CS46)	63	18	81	P
Design and Analysis of Algorithms Laboratory (10CSL47)	37	18	55	P
Microprocessors Laboratory (10CSL48)	24	18	42	P

Semester: 3 Result: FAIL

August 26, 2015

RESULTS (PROVISIONAL)

Results are not yet available for this university seat number or it might not be a valid university seat number

Key implementation issues -5

- **Scraping without getting noticed**
  - Scraping may not be encouraged if we bring the website down by relentless query
  - How can we impersonate random user ?
    - incorporate random delay in between HTTP request
    - Choose a scraping framework that simulates a browser
  - How can we minimise no of website hits ?
    - Do not scrape for 1-999 for 153 colleges. it may bring the website down or my IP will be blocked or a police complaint will be raised !
    - Total no of students “x” but maximum USN can be “x+y” !
    - May manually check FastVTUresults.com to see
      - Total no of students in each college
      - Ending USN at each college
      - List of colleges offering the course

Key implementation issues -6

- **Choose a database**
  - HTML result format itself is inconsistent
    - subject 1 in one may be subject 3 in another (subject itself a variable !)
    - For some students, we get blank or partial result
    - Students who cleared "backlog from previous semester" has a different HTML result page
    - Hard to fit into a RDBMS like MySQL
  - Result is a document where subjects, college/student details are subdocuments. So, a NoSQL type JSON database such as MongoDB is a good fit!
    - MongoDB specific tools need to be learnt
      - Writing into MongoDb using Python Pymongo framework
      - MongoVue and RoboMongo as GUI tool for MongoDb
      - MongoDB syntax and aggregation frarmework

Key implementation issues -7

- **Choose data analytics platforms**
  - MongoDB has no open source data analytics package.
  - MongoDB aggregation framework is limited
  - Research and evaluate available tools for data analytics with NoSQL databases
  - Solution
    - Use Pentaho Data Integration tool to export Mongo Data to Weka Data Mining workbench
    - Use Weka Data mining workbench for statistical analysis

Key implementation issues -8

JSON document schema

Robomongo 0.8.5

File View Options Window Help

mongodb (4)
- System
- testdb
- vtu
  - Collections (3)
    - System
    - SCS02
    - SCSRGIT
  - Functions
  - Users
- yelp

db.getCollection('SCS02').find({})

mongodb localhost:27017 vtu

```
db.getCollection('SCS02').find({})
```

SCS02 0.002 sec.


```
        "tot" : 46
    },
    "14SCS22" : {
        "int" : 48,
        "ext" : 0,
        "passfail" : "A",
        "sub" : "Advances in Computer Networks (14SCS22)",
        "tot" : 48
    },
    "college" : {
        "collegeCode" : "1AK",
        "collegeName" : [
            "AKSHAYA INSTITUTE OF TECHNOLOGY"
        ]
    },
    "14SCS24" : {
        "int" : 48,
        "ext" : 0,
        "passfail" : "A",
        "sub" : "Artificial Intelligence and Agent Technology (14SCS24)",
        "tot" : 48
    },
    "student" : {
        "nameusn" : "KANTHEESH S (1AK14SCS01) ",
        "sem" : "2",
        "result" : " Result:  FAIL "
    },
    "14SCS26" : {
```


RoboMongo GUI

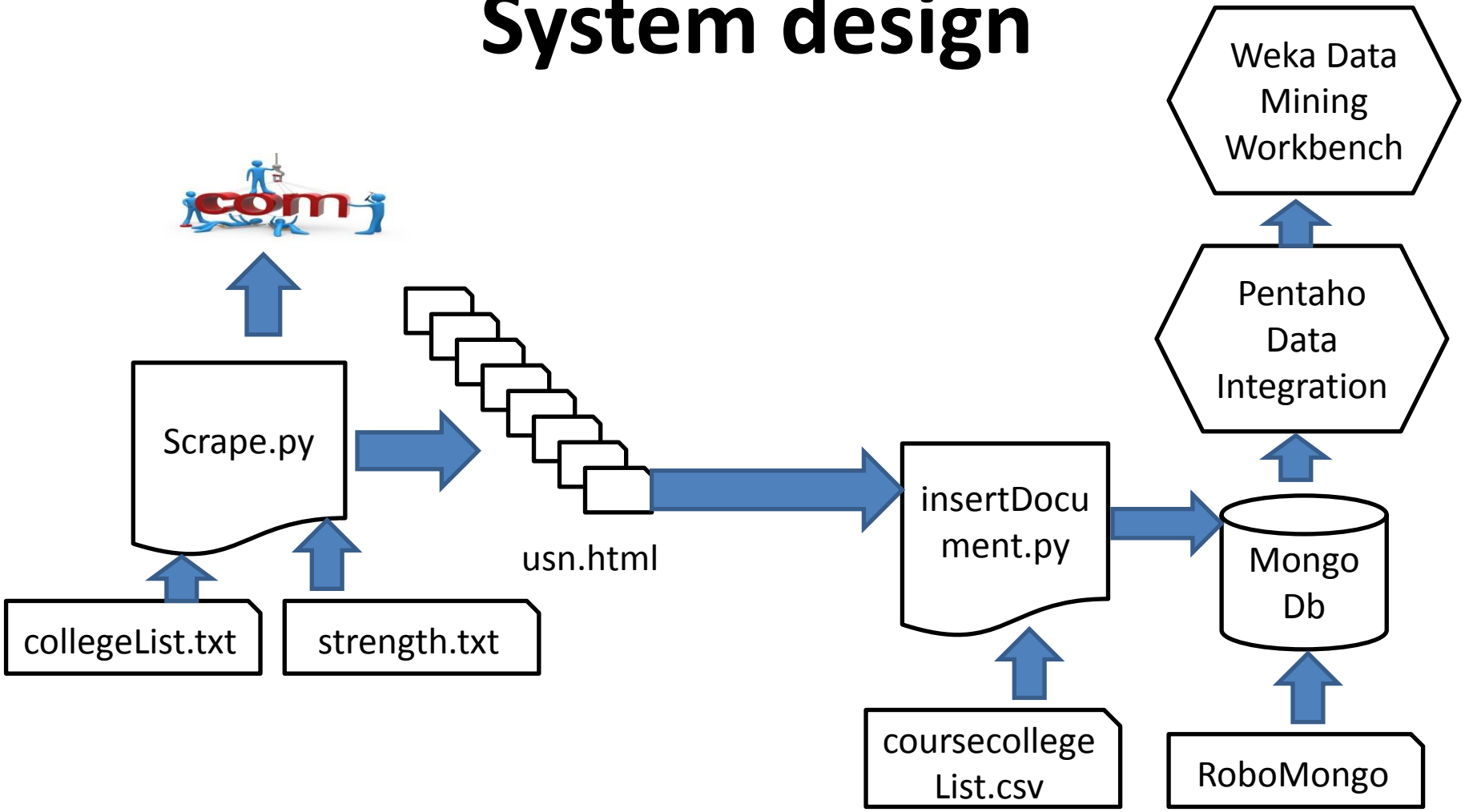
Key implementation issues -9

- **Long debugging cycle**
  - For each semester dataset (while forming the JSON document)
    - Typically, if a program fails, one looks at the logic of the code
    - Here, we look at the result (USN.HTML) where the code has failed to understand the reason
    - We do this many times for every dataset !

Summary of steps

- **Step 1 :** We make **a list of VTU colleges offering the course** and for that particular semester, note down the no of students and maximum USN number from http://fastvturesults.com
- **Step 2 :** Get the **HTML result page from VTU Results website** (http://results.vtu.ac.in/) by automatically filling in form (student USN). We do this with some random delay in between HTTP requests . Write and test the code for this.
- **Step 3: Scrape the HTML result web page to get the subjects and the score** (internal, external, total, final result) and student details (name, USN, result) data. Write and test the code to do this.
- **Step 4 : store the data in MongoDB** (noSQL database)
- **Step 5 : export the Mongo DB data** (using standard data pre-processing techniques) using Pentaho Data Integration (Community Edition) to a .ARFF file for importing **into Weka Data mining workbench**
- **Step 6 : Use Weka Data mining workbench** (open source) to do statistical analysis
- **Step 7 : repeat step 2-6 for RGIT** ONLY to do the comparison

System design



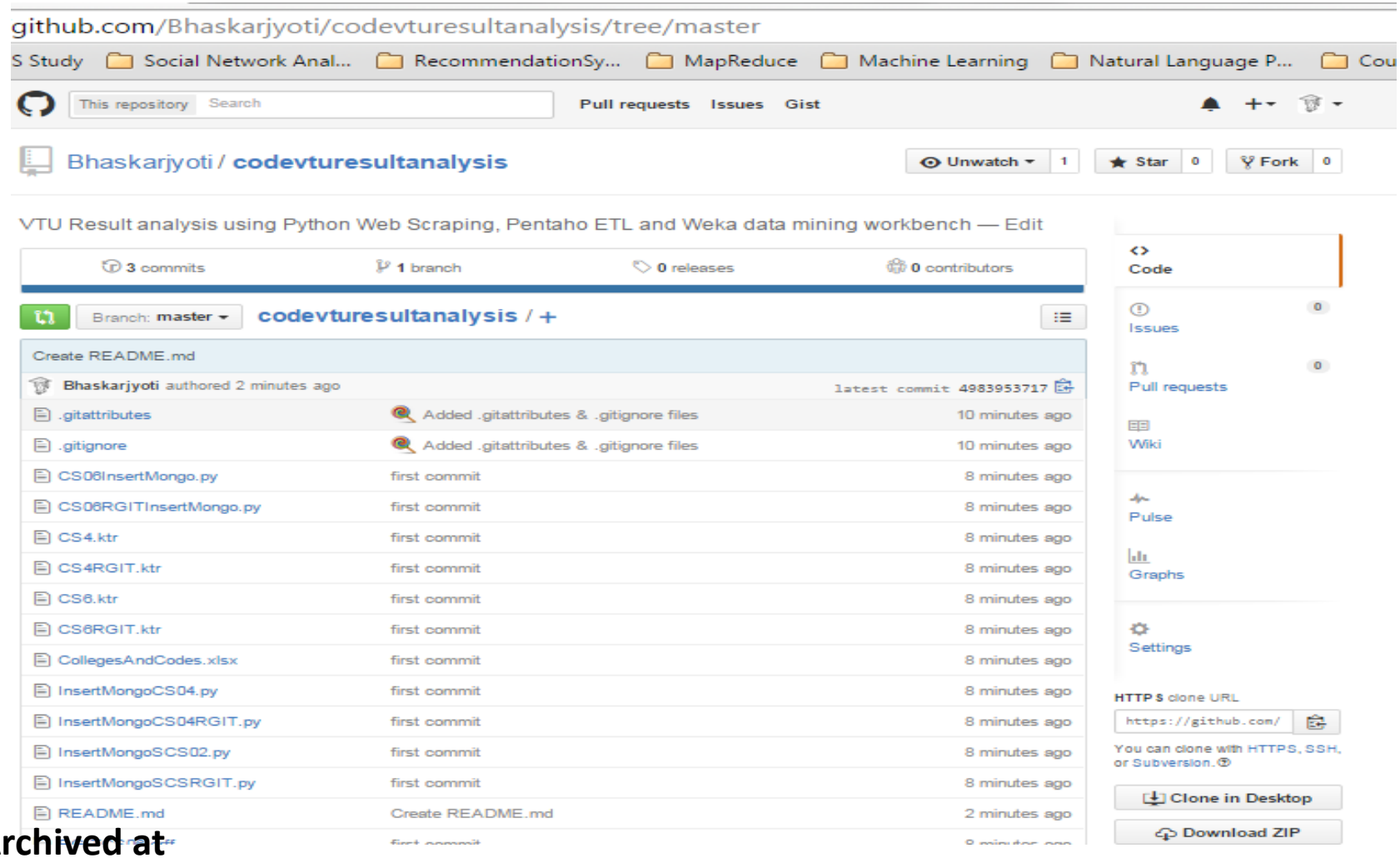
Future improvements

- A program can be developed to create a per college (python) list (of USN) by web scraping fastvturesults.com and use this list to scrape the USN
  - This may not be easily possible as this requires multiple buttons to be selected in more than one web page
  - the existing scraping frameworks may not allow this

Learning outcome

- Technology
  - Python scraping frameworks
  - NoSQL database (MongoDB)
  - ETL tool (Pentaho Data Integration)
  - Data mining workbench (Weka)
- Introduction to real life data mining problem
  - Data based debugging of code
  - Acquiring and cleaning dataset is 80% of the effort!
- Delivered what was asked for !

The code



Archived at

https://github.com/Bhaskarjyoti/codevturesultanalysis.git