Assignment 2: Team Based Development – Study 1

SENG 403, Winter 2016

Team T02-2

# 3-1) Design Steps Taken:

SQL queries and data parsing:

INSERT STEPS

Calculate Median Tenureship and Tenure Diversity:

INSERT STEPS

Inferring Gender using Gender Computer and Python script genderParse.py:

In the study GenderComputer is a python library that infers gender based on a given real name or user name and a country. This was used as a tool to help infer gender in GitHub users. The problem we faced was the fact that much of the data in the ghTorrent Dump had cities, or states, or a combination of the two – with no country being shown. This was a problem as GenderComputer did not handle cities or states. Our initial solution was to get a database or a dictionary in python that maps cities to country, and this was unsuccessful. The task of finding a solution was given to Ken, in which a script was made to accurately find a country associated with a city. The solution involved taking 3 additional libraries:

BeautifulSoup - <http://www.crummy.com/software/BeautifulSoup/>

Openpyxl -https://pypi.python.org/pypi/openpyxl/

Wikipedia library for python - <https://pypi.python.org/pypi/wikipedia>

The actual script itself will be attached with this file, but the main process is this:

1. The script extracts the location for a given user using openpyxl
2. It then tries to parse any country in it by using a dictionary that contains all the countries in the world in it. If something is found it uses the country and user, passes it to gender computer in which it tries to infer the gender and places the gender within the excel sheet.
3. If nothing is found it then sends a search request to Wikipedia using the Wikipedia api
4. It will then take a result found an get the HTML page from it
5. There is a constant pattern in how Wikipedia sets up their city pages in which the right column will house the data for Country associated with it.
6. Using BeautifulSoup, it will parse the HTML to find Country within that column, and when it is found we know that the actual country name will be the sibling to the element in which Country is found.
7. We then extract that country and pass it to gender computer in which it will infer gender.

There are some sources of error in this – if any of the city HTML pages don’t follow that pattern then a country may not be found, Multiple cities of the same name (It may send back the wrong Country associated with the city). But these kind of problems did not show up often.

Rudimentary Analysis using Scatterplots and placing trend lines to observe the trend of the data:

INSERT STEPS

# 3-1) List of Tasks

Do we just rewrite what is in GitHub?

# 3-3)Re-run of the Study

## Results:

H1:

H2:

H3:

H4:

## Comparisons on findings:

# 3-4)Description of code and analysis artifacts

# 3-5) Future Work