Letter to future DIY researchers:

I guess this is where I document stuff for you. I’ll try to break up my documentation into pieces, so all the information doesn’t run together.

Goal:

* Find a correlation between types of bacteria in people’s saliva and their health/habits/characteristics

Current Stage:

* Some bacteria classes have means that are significant for certain subcategories with a 95% confidence
* Machine learning based off of those significant bacteria classes yield testing accuracies for categories to as high as 90% accuracy rate to as low as 15% accuracy rate

Problems:

* Current problem is the lack of data. **This needs to be addressed first before you move on with this project.**
  + Machine learning training and testing sets range around 40-52 testable sets of data (so results are super unreliable)
* The FRIome project has been going on since 2014, so we have test data from all the way back then. However, the data in the database only contains 2 years worth of data (2014-2015 and 2015-2016). **Data for the more recent years exists** on UTBox but in fastq form (so a bunch of DNA sequences), so they need to be analyzed and uploaded into DIY’s database (phpMyAdmin).
  + If you want to look at the data, have Tim grant you access to UTBox and phpMyAdmin
  + Raw data should be found in one of the JA----- folders (UTBox). To figure out which one, take a look at the job names from here: [documentation](https://mail-attachment.googleusercontent.com/attachment/u/1/?ui=2&ik=39772f763b&view=att&th=165a5e6a6774b90a&attid=0.4&disp=inline&realattid=f_jlo2b4w03&safe=1&zw&saddbat=ANGjdJ9NfMNcc06Ofr8UAkhwvqKQuTXfzbha5YkMURME40uYrnSupaeURnvgz4dNHuJv6wpDD2JAG3up3E-_-hIynQtqv3rd_HqmuAOWyIMrnQlDih-1lqQ9RTerJcv7z3dSnVYh0Vp8w10IczJsp6sbjIgF1A69hGeY2iiajmSWv0snDfUXbKt-0F0FuTt872pB3oDVeg6exh4urDEBjvlodqkFckMQv1txetkJUjOYYVOUMC9m31Bthjxswj9X_6BxUbDnSBqlsPB4a_plS1Aao3tO9oBaE7ney7fCq-r6-akpFjCkgDBA8vFrCWfzIl4bpiUFJqt0eXg5MENldB0vC-LNY2LnujYll-DqFvynHeaFALyFyrVW5vL4Dfd3aQkvdT7x-XqmBYvccQfcw1wHbpysUl5xdd_I1gysIQRpDDDE938g4RC3OGJFmOeQro9Q15d-l7sfHxMY1jox2BbpAkpH9Yz_e7cvvImoub_fZ21xD45J6XJk5tZTf1tHHwR7BCBCVABppgFM8Ph42E7BBN2ULRyjoB3CjKxfXv3GqRbZJoICCqG1lK43txM01bqXFS03zBBL8xtStDDmQhN_EEuTFa00a6M95MvRj3C7JtiTq-pHaaXAh80jetM2i0rv36VuFm-7rM4UmOxF)
  + phpMyAdmin has a database called FRIome. You can kinda ignore Table13 since I created that and didn’t know how to rename it, but I think it’s either the combination of bacteria class data from the two years or the bacteria class data from 15-16

Other problems I ran into:

* The survey form for 2014-2015 is structured differently than the form for the more recent years (so 1, 2, 3, 4, 5 may mean bad, poor, okay, fine, good, etc. or different columns represent different questions). I tried to map/fix some of results from 14-15 to results from 15-16 in my helper methods near the top of my code.
* Not everyone answered all the questions. Some people misspelled stuff, so that makes it difficult parsing. I also addressed this in the helper methods.
* Some ID’s from the data doesn’t match the ID’s from the survey
* There are multiple sample results per ID. This is not necessarily a problem, but rather it takes time to decide which one you want to use.
* Some categories shouldn’t be tested on due to how skewed the results are (ex: smoking, neardogsweek, nearcatsweek) and would therefore provide skewed results. The fix for this would be more data collected.

Steps I took for this project:

* Figure out which survey categories I wanted to find correlations for
* Figure out which taxonomic rank for the bacteria I wanted to find correlations for
* Figure out which samples I wanted to use (since most ID’s had multiple samples to choose from, and some ID’s didn’t have enough data or didn’t correlate with the ID’s from the survey responses)
* Once I had all of the above information, I created a SQL query which gathered all the above information from the database (each ID I wanted to use and their corresponding survey results and bacteria count). This is so that it will output a master spreadsheet which I can automate a python script on.
* Afterwards, I deleted a couple of the classes at the end that have few info. They might screw up results. I also normalize the info for the rest of the classes. This was done in Excel.
* Python script automates the statistical testing of the data. Within each category, it compares each bacteria class with itself for each subcategory.
  + For example, in the category gender: compare class A in males to class A in females and see if there is a significant difference between the two means
  + T-tests are ran for categories with only 2 subcategories, which ANOVA tests are ran for categories with more than 2 subcategories.
* Once the significant classes for found for each category, use them as input nodes to create a neural network for each category to see if we can predict the subcategories given those classes
  + Total data is split into 2: testing and training
  + Results would differ for each category every time the script is ran (as a different neural network is created every round). But results for statistical testing should stay the same.
* Software/Services and Languages/Libraries used:
  + Software/Services: phpMyAdmin (database), Microsoft Excel, Jupyter Notebook (for Python script), Git/Github (for storing files and version control)
  + Languages: MySQL (just a little bit in order to manipulate database), Python
  + Python Libraries:
    - Enum, pandas, scipy, stats, numpy (for general math/data structures)
    - Matplotlib (for graphs)
    - Tensorflow, keras (for machine learning)
    - Os (for converting notebook to html)
* My files:
  + My Github is kinda a mess old files that I used once and scrapped. Only around 3ish files matter: <https://github.com/madelinehuang/FRIome>
  + Main files:
    - Jupyter Notebooks/Ttest-Anova-Calc.ipynb
      * This is where all my code is
    - Jupyter Notebooks/MasterSheet.csv
      * Master spreadsheet that contains the data inputted into Ttest-Anova-Calc
    - Jupyter Notebooks/MasterSheetKey.csv
      * Key for the categories
  + Other Notebooks (probably won’t run since input files are in other folders):
    - PhylumCorrelation is my very first file which calculates correlation matrices for 1 category (gender) on bacteria phylums
    - Class-GraphingComparison graphs and compares bacteria classes and subcategories for 1 category (gender)
  + Filtering Data - Some highlighted files while I was deciding which samples/categories to use
  + Original files - data downloaded/selected from the database
  + BacteriaInfoDump.xlsx - starting file when I was trying to get a big picture of the data
* Final thoughts:
  + I’m pretty sure there are other ways to approach this project. Nina Hua a couple of years ago used Rstudio to calculate correlation. I mainly stuck to Jupyter Notebook because I liked coding in Python and wanted to maybe do some machine learning.

Good luck and have fun,

Madeline Huang

Dec 2018