MAYO INTERNATIONAL SCHOOL

Patparganj, Delhi

CENTRAL BOARD OF SECONDARY EDUCATION
Committed to Equity and Excellence in Education

INFORMATICS PRACTICES
AISSCE
PROJECT SYNOPSIS

2021-22

EMAIL DATA VISUALIZATION AND ANALYSIS

HEAD OF DEPARTMENT:
MS. SAPNA RAI

SUBMITTED TO DEPARTMENT OF INFORMATICS PRACTICES FOR THE PARTIAL FULFILLMENT OF AISSCE EXAMINATION SESSION – 2021-22.

ACKNOWLEDGEMENT

In performing our assignment, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. The completion of this assignment gives us much Pleasure. We would like to show our gratitude to Ms. SAPNA RAI, Course Instructor, for giving us a good guideline for assignment throughout numerous consultations. Team member have made valuable comment suggestions on this proposal which gave us an inspiration to improve our project. We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us in writing this assignment.

With sincere thanks.

CERTIFICATE

This is to certify that Mr. Aashish Raj, Mr. Achal Jain, Ms. Kuhoo Gangwar, student's of Class XII E, completed the Term I – PROJECT SYNOPSIS "Email Data Visualization and Analysis", during the academic year 2021-22 towards partial fulfillment of credit for the Informatics Practices AISSE Project work of CBSE and submitted satisfactory report, as compiled in the following pages, under my supervision.

DATE:

Head of Department Signature:	Principal Seal and Signature:

PROJECT LOGBOOK

PROJECT NAME: Email Data Visualization and Analysis.

YEAR: 2021-22.

CLASS: 12th E

TEACHER NAME: Ms. Sapna Rai.

TEAM MEMBER NAMES

1. Achal Jain.
2. Aashish Raj.
3. Kuhoo Gangwar.

INDEX

INTRODUCTION

Reading, responding to and even organizing emails can oftentimes end up being an enormous time sink. While the numerous email clients out there today go to great lengths to make life easier in this regard, it could be really useful if one could perform email-related tasks in a programmatic manner. Let us consider a scenario where this approach can come in handy — email clutter. Now, it can be argued that the internet is awash with advice on how to reduce email clutter. While all these tips can potentially help you use your email more efficiently going forward, what if currently, your inbox has reached a point of no return, where it is nearly impossible to clean things up manually? A classic case in point — my inbox has over 35000 emails from over 300 different senders. I am positive that most of these are marketing related, and it feels, and rightly so, that it is way too much of an effort to individually unsubscribe to these emails. The first part of this project deals with this issue and explains how to send, retrieve, categorize, delete and unsubscribe to emails using python's imapclient and smtplib libraries.

On the other hand, let us assume you are an individual who is in total control of your inbox and meticulously labels and files every single message. You might be interested in some statistics about your emails like who emails you the most, what does your traffic look like or what your typical email response times are. The second part of this project is meant exactly for this, demonstrating how to analyze email data using Pandas and create visualization tools using Matplotlib.

TEAM ROLES

ROLES OF TEAM MEMBERS:

ROLE	ROLE DESCRIPTION	MEMBER NAME
★ To retrieve data from an Email account after a user entered input providing Mail ID and password ★ To parse the file and create a CSV file of mail data ★ To import the file into python and read it as a Data Frame ★ To view a certain mail on demand based on parameters received by use	To Visualize To Organize To Manipulate	ACHAL JAIN
★ To perform operations on mails and also find common words used ★ To delete mails on demand after receiving parameters from user ★ To run a program on demand to delete certain categories of mails (eg spam mails)	To Analyze To Manipulate	ASHISH RAJ
★ To create various graphs using matplotlib such as weekly email traffic, hourly email traffic, Top email senders, subject word count, common words in subject headers ★ To check, categorize and create sub DataFrames based on their Type e.g., spam, Promotions, Important etc. This will be done via a list of pre-saved words to be search for in mail subjects + taking input from user	To Analyze To Visualize	KUHOO GANGWAR

PROJECT PLAN

PHASE	TASK	ACTUAL START DATE	ACTUAL END DATE	RESPONSIBILITY
PREPARING FOR THE PROJECT	COURSE-WORK READINGS	15 September 2021	20 september 2021	Achal
	SETTING UP TEAM FOLDER	15 September 2021	20 september 2021	Achal
DEFINING THE PROBLEM	BACKGROUND READING	15 September 2021	20 september 2021	Achal
	TEAM MEETING FOR TOPIC SELECTION	15 September 2021	20 september 2021	Kuhoo
BRAINSTORMING	TEAM MEETINGS FOR IDEA GENERATION	15 September 2021	20 september 2021	Kuhoo
DESIGNING YOUR SOLUTION	TEAM MEETINGS TO DESIGN THE SOLUTION	15 September 2021	20 september 2021	Aashish
COLLECTING AND PREPARING DATA	TEAM MEETING TO DISCUSS DATA REQUIREMENTS	15 September 2021	20 september 2021	Kuhoo
PROTOTYPING	DATA COLLECTION	15 September 2021	20 september 2021	Achal
	PREPARATION AND LABELLING	15 September 2021	20 september 2021	Kuhoo, Aashish
PROTOTYPE TESTING	PERFORM DESIRED OPERATIONS	15 September 2021	20 september 2021	Aashish
	INITIATE ACTIONS BASED IN THE RESULT OF YOUR MODEL	15 September 2021	20 september 2021	Aashish

Communications plan

1. Will you meet face-to-face, online or a mixture of each to communicate?
   Ans. We will meet online for all communication

2. How often will you come together to share your progress?
   Ans. We came together once a month

3. Who will set up online documents and ensure that everyone is contributing?
   Ans. We took collective responsibility

4. What tools will you use for communication?
   Ans. Google Meets

#1 Team meeting minutes

Date of meeting: 8/7/2021
Who attended: All Members
Who wasn't able to attend: N/A
Purpose of meeting: To discuss to the topic and project details
Items discussed:

1. What is our Topic ?
2. Work Distribution
3. When to work

Things to do (what, by whom, by when)

1. Research - By All - 1 month time
2. Concept Practice - By All - 1 month time

#2 Team meeting minutes

Date of meeting: 25/10/2021
Who attended: All Members
Who wasn't able to attend: N/A
Purpose of meeting: To discuss to the topic and create synopsis

Items discussed:
1. Topics Details and Subtopics
2. Synopsis Creation

Things to do (what, by whom, by when)
1. Complete Assigned Work - By All - 1 month time.

#3 Team meeting minutes

Date of meeting: 25/10/2021
Who attended: All Members
Who wasn't able to attend: N/A
Purpose of meeting: To compile the data

Items discussed:
1. To submit individual content and compile data
2. To Collect all Data and finishing touches

Things to do (what, by whom, by when)
1. Take print and submit - by Aashish - when asked to submit.

PROBLEM DEFINITION

1.Important issues faced by school/community.

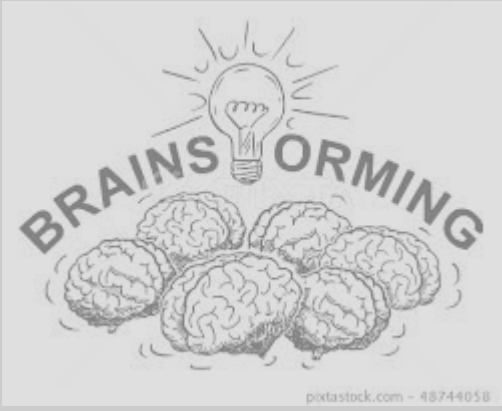
The issue we wish to deal with is the management of mails we receive everyday. We get mails from different people and of different kinds. We wish to find out the frequency and details of such mails to help align and better organize our mails. We often even get spam mails and promotion mails. We need to find which companies/organizations promotional mails the most and consequently deal with them

2.Which issues matter to you and why?

The issues which affect us most are

1. The need to organise our mails
2. The need to detect promotional mails and advertisement
3. The need to segregate between important mails and spam mails

BRAINSTORMING



4.1 Ideas

How might you use the power of Data Science using PANDAS and MATPLOTLIB to solve the users' problem by increasing their knowledge or improving their skills?

Idea #1	By segregating the data into subgroups so that it is easy to access.
Idea #2	Creating a program so that the spam folders will be removed efficiently.
Idea #3	By visualising the data into different graphs to provide a clear picture of what kind of data is possessed by the user.
Idea #4	By performing operations on mails and also find common words used
Idea #5	By deleting the mails on demand after receiving parameters from user

4.2 Briefly summarize the idea for your solution in a few sentences and be sure to identify the tool that you will use.

- Summary - We plan to analyse and visualise the emails in such a way by which the user can have access ,control and most importantly a clear cut image of the source of the data.

- Tools Used - 1. Python Pandas
  2. Matplotlib.pyplot
  3. Numpy
  4. Openpyxl
  5. Getpass

Data

5.1 What data will you need for your project?

CSV of emails to manipulate,analyse, and visualise the data.

5.2 Where or how will you source your data?

Data needed	Where will the data come from?	Who owns the data?	Do you have permission to use the data?	Ethical considerations
Have	Email IDs of group members	Group Members	YES	The group members have given allowance for use
Want/Need	Emails of group members	Group Members	Yes	The group members have given allowance for use
Nice to have	Emails of a random user	User	Depends on User	The user needs to give allowance to use her/her mails

PROTOTYPE

6.1.Tools used to build the prototype.

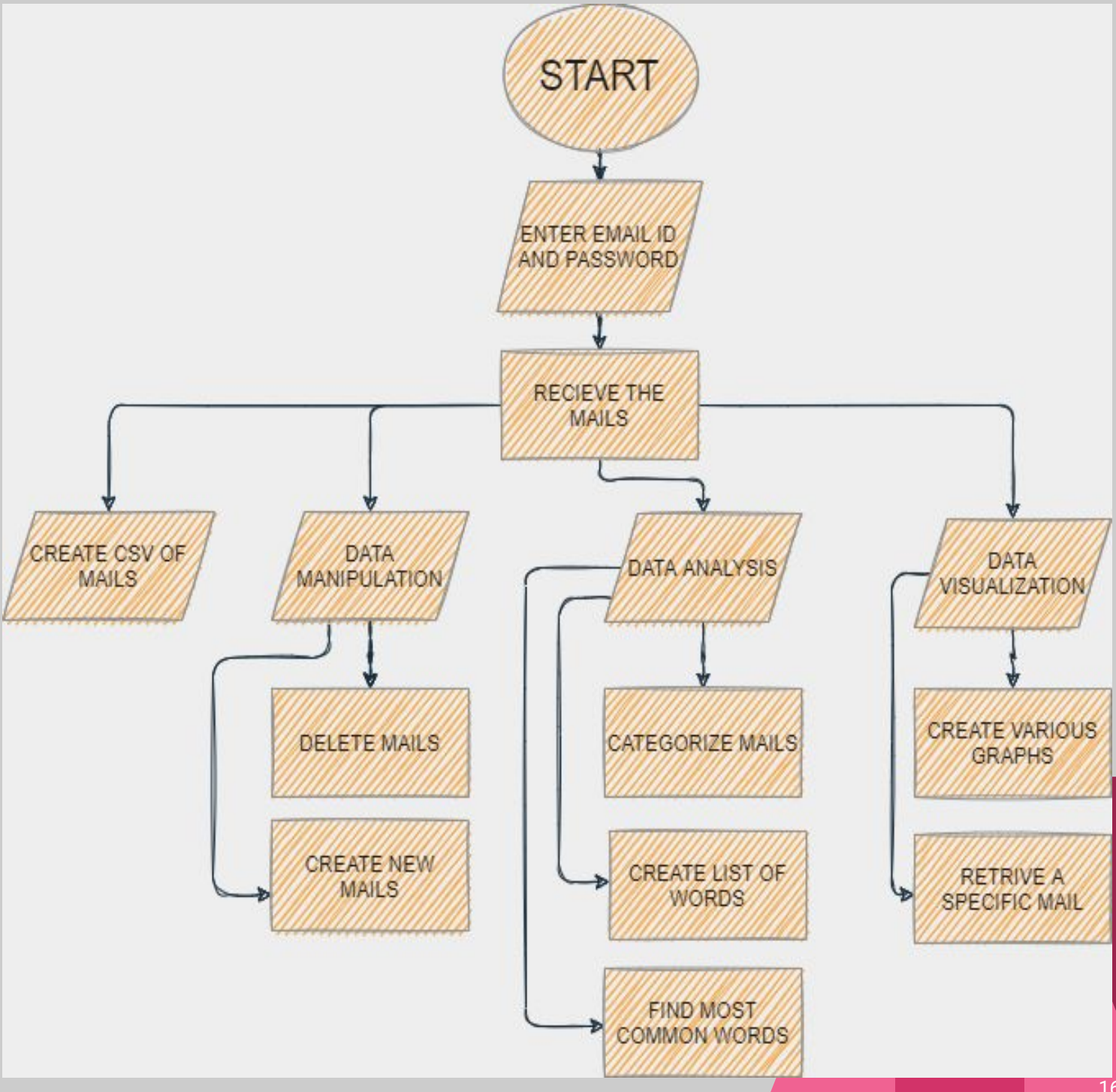
- Python Pandas
- Matplotlib.pyplot
- Numpy
- Openpyxl
- Getpass

6.2 What decisions or outputs will your tool generate and what further action needs to be taken after a decision is made?

- We will be able to divide the data into different sub - groups.
- We will be able to delete any desired mails based on the user's demands.
- We will be able to filter the spam mails and save the memory and storage
- We will be able to visualise the data into different graphs.
- We will be able to run a program on demand to delete certain categories of mails
- We will be able to change data of a mail to insinuate different values and then re-analyze
- We will be able to perform operations on mails and also find common words used
- We will be able to view a certain mail on demand based on parameters received by user.

- FURTHER ACTIONS:: To run the program and check for any gap or error in it.

Methodology/Flow Diagram of the proposed work



Hardware and Software Used

- **PROCESSOR:**

  Intel(R) Core(TM) i5-1035G1

  CPU @ 1.00GHz 1.19 GHz

- **RAM**

  8.00 GB

- **OPERATING SYSTEM**

  Windows 10

- **SYSTEM TYPE**

  64-bit operating system, x64-based processor

- **PYTHON VERSION**

  Python 3.7.8

Python Module Used

- **PANDAS**

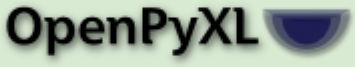
In computer programming, pandas is a software library written for the Python programming language for Pandas is an open-source library that is made mainly for data manipulation and analysis of data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on the top of the NumPy library.Pandas is fast and it has high-performance & productivity for users.

- **MATPLOTLIB**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Python Module Used

- **OPENPYXL**



Openpyxl is a Python library for reading and writing Excel (with extension xlsx/xlsm/xltx/xltm) files. The openpyxl module allows Python program to read and modify Excel files. Using Openpyxl module, these tasks can be done very efficiently and easily

- **NUMPY**

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.NumPy is open-source software and has many contributors.

- **GETPASS**

getpass() prompts the user for a password without echoing. The getpass module provides a secure way to handle the password prompts where programs interact with the users via the terminal.

Analysis/Manipulation/ Visualisation Work Description

To Visualize:

1. To retrieve data from an Email account after a user entered input providing Mail ID and password
2. To parse the file and create a CSV file of mail data
3. To import the file into python and read it as a Data Frame
4. To view a certain mail on demand based on parameters received by user
5. To create various graphs using matplotlib such as weekly email traffic, hourly email traffic, Top email senders, subject word count, common words in subject headers

To Analyze:

1. To check, categorize and create sub Data Frames based on their Type e.g., spam, Promotions, Important etc. This will be done via a list of pre-saved words to be search for in mail subjects + taking input from user
2. To perform operations on mails and also find common words used

To Manipulate:

1. To delete mails on demand after receiving parameters from user
2. To run a program on demand to delete certain categories of mails (eg spam mails)
3. To create mails/new input in the DataFrame to insinuate a received mail
4. To change data of a mail to insinuate different values and then re-analyze
5. Back to STEP 1 TO REANALYZE

List of References

- https://medium.com/analytics-vidhya/how-to-read-and-write-data-to-google-spreadsheet-using-python-ebf54d51a72c
- https://towardsdatascience.com/email-automation-analytics-and-visualization-53b022e0f9a0
- http://beneathdata.com/how-to/email-behavior-analysis/