CSE 508
Divyansh Upreti (112026646)
Design Challenge Submission

Spam Reporter

Spams - 'ubiquitous, unavoidable and repetitive'.

Have you ever faced a problem when you visited your spam box bombarded by tons of emails and wondered if you knew which spam emails you needed to delete and which spammers you needed to unsubscribe ? Have you ever wanted to track down the spammer who spams you the most. Have you ever wondered in which area/field you receive the most spam emails ? To know these answers easily, I have made a program in python which tells the the most common word received in a person's spam mail box, provides graphical analysis of emails from top 10 spammers and tells location, ip address and the count of the emails sent by each spammer. This can help anyone to get idea about which emails he needs to unsubscribe on a priority basis and what are his interests and areas of focus.

Email spam has steadily grown since the early 1990s. Botnets, networks of virus-infected computers, are used to send about 80% of spam. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. This makes it an excellent example of a negative externality.

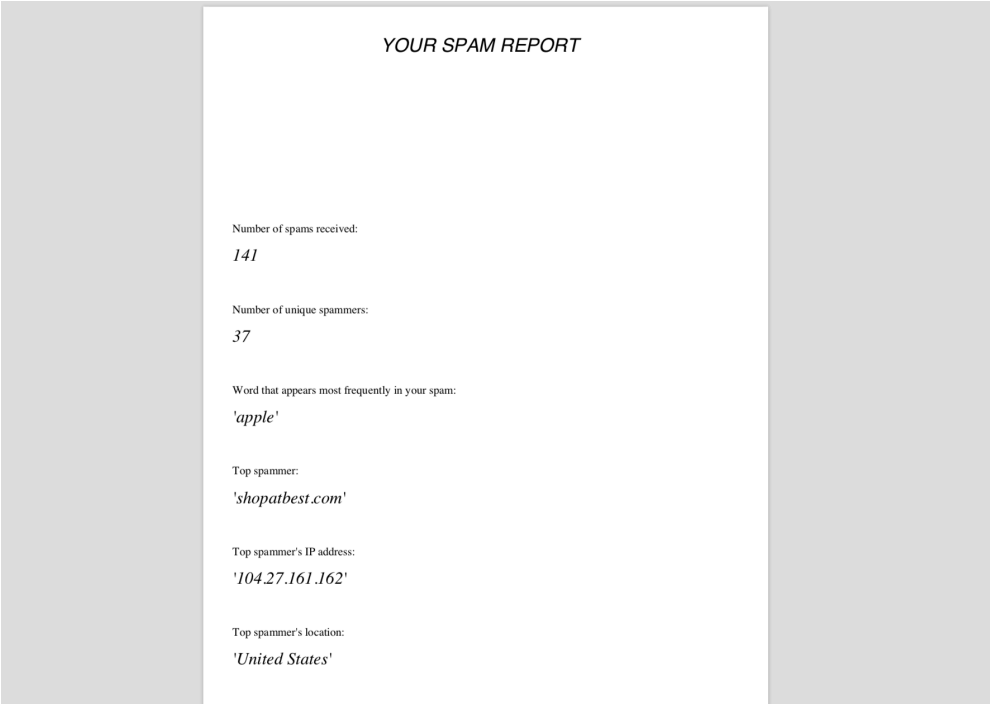
Keeping in mind the difficulties we face due to spam, the program contained in this project can help mitigate some of our problems by providing necessary information. In the rest of the report I have presented the information It provides and the steps needed to provide spam content to the program and how to interpret results.

What all this program does:

This program opens a PDF file ‘result.pdf’ when it finishes executing that displays the following information:

1. ***Page 1 :SPAM BOX SUMMARY*** -

   It tells the number of spam emails present in the spam folder, number of unique spammers and gives the location and domain information of top spammer.



YOUR SPAM REPORT

Number of spams received:
141

Number of unique spammers:
37

Word that appears most frequently in your spam:
'apple'

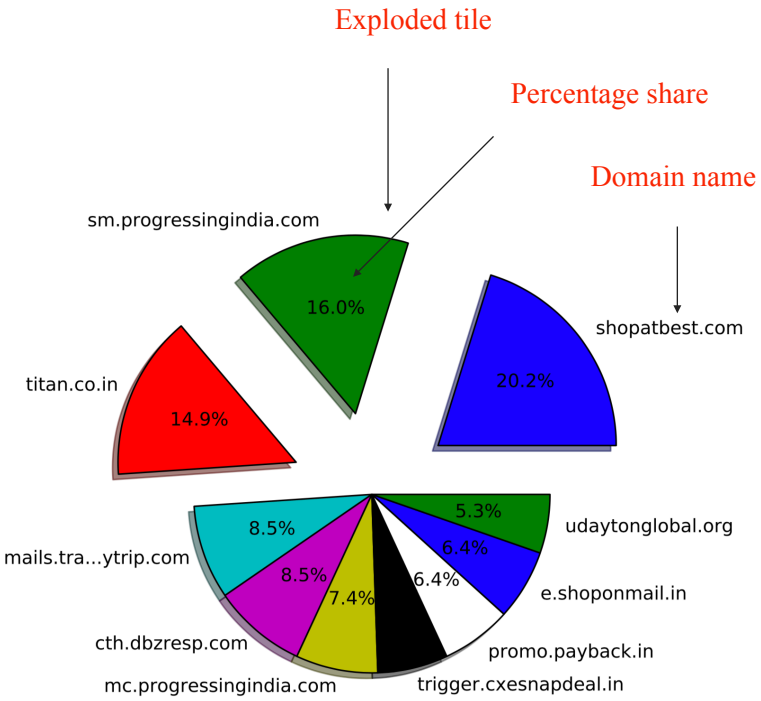
Top spammer:
'shopatbest.com'

Top spammer's IP address:
'104.27.161.162'

Top spammer's location:
'United States'

2. ***Page 2: GRAPHICAL ANALYSIS OF TOP 10 SPAMMERS-***

The graph displays top 10 spammers and tells percentage share of each spammer in spam box. Top three spammers are distinguished by exploded pies in the pie chart as shown:



3. ***Page 3 & ahead: COMPLETE DATA-***

Presents the data about each and every spammer present in the spam mail box in the form of a table. The information that we can get from the table is the domain name of spammer, correct ip address and location and the number of emails sent by the spammer. If the IP address/location is unavailable, 'Unavailable' is shown in the cell.

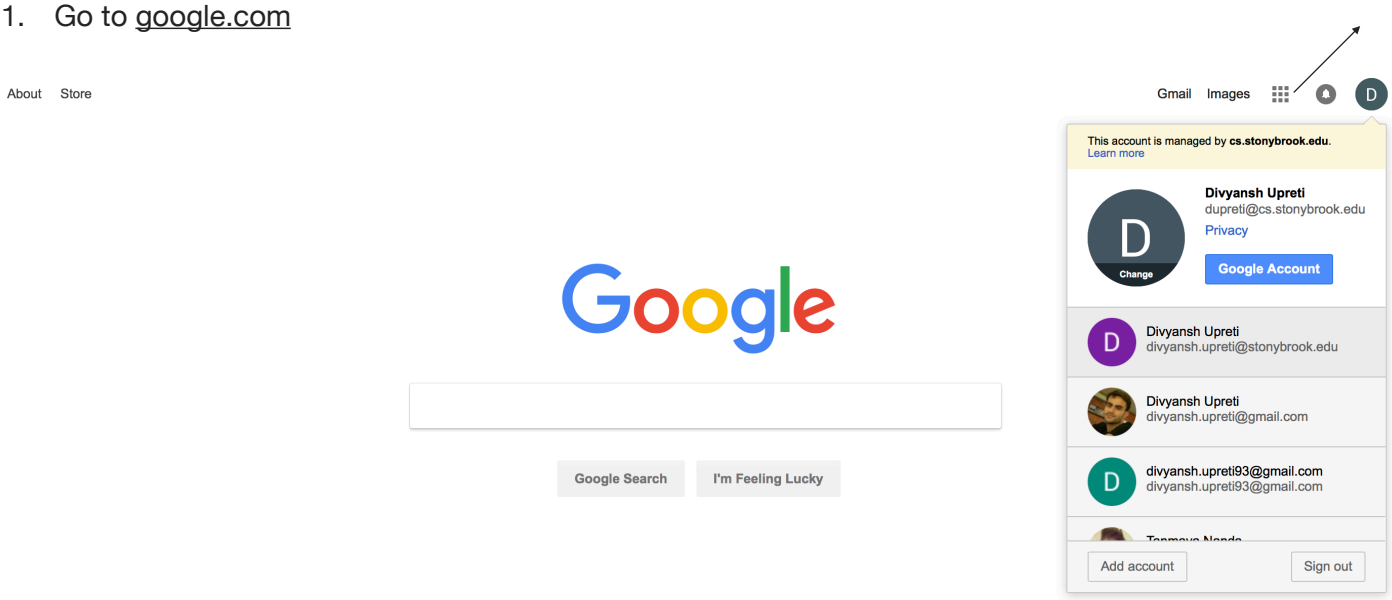
Rank	Count	Domain	IP Address	Location
22	1	srv.jobs-on-mail.com	115.166.137.43	India
23	1	shriramgeneral.co.in	50.63.202.8	Scottsdale, United States
24	1	referhire.com	35.154.0.41	Mumbai, India
25	1	rajcomics.com	52.2.107.9	Ashburn, United States
26	1	mta.updates-on-mail.com	115.166.137.54	India
27	1	mobilegcase.com	166.62.6.67	Scottsdale, United States
28	1	lykapp.com	34.226.89.241	Ashburn, United States
29	1	kitabaystore.in	50.63.202.40	Scottsdale, United States
30	1	email03-c...iever.net	67.227.173.191	Lansing, United States
31	1	em.soda-pdf.com	192.196.223.6	Canada
32	1	digitaladdis.com	67.209.238.3	Scottsdale, United States
33	1	customers...email.com	Unavailable	United States
34	1	codechef.com	54.226.138.241	Ashburn, United States
35	1	blk.mouthshut.com	64.237.46.235	Matawan, United States
36	1	beta.glob...rends.com	115.166.137.65	India
37	1	amrita.edu	103.10.24.196	India

Rank	Count	Domain	IP Address	Location
2	15	sm.progressingindia.com	216.117.184.145	Fayetteville, United States
3	14	titan.co.in	23.32.175.153	Cambridge, United States
4	8	mails.tra...ytrip.com	Unavailable	Fayetteville, United States
5	8	cth.dbzresp.com	134.119.195.171	France
6	7	mc.progressingindia.com	216.117.143.36	Fayetteville, United States
7	6	trigger.cxesnapdeal.in	Unavailable	Lansing, United States
8	6	promo.payback.in	8.33.184.254	United States
9	6	e.shoponmail.in	63.149.195.18	Tustin, United States
10	5	udaytonglobal.org	Unavailable	India
11	5	imaginestore.org	50.87.150.247	Provo, United States
12	4	crm.cxesnapdeal.in	Unavailable	Mumbai, India
13	4		0.0.0.0	Unavailable
14	3	naturalhealthsherpa.com	72.52.188.98	Lansing, United States
15	3	email05-c...iever.net	67.227.173.191	Lansing, United States
16	3	amcatmail.com	184.168.221.63	Scottsdale, United States
17	2	lemontreehotels.com	103.11.86.84	Noida, India
18	2	jobs.geebo.com	188.166.204.77	Singapore, Singapore
19	2	irctc.co.in	103.252.142.19	India
20	2	edm.efinmail.com	67.134.222.254	Sheffield, United States

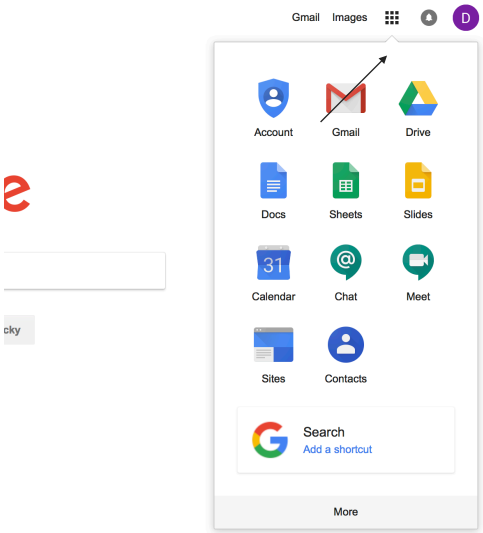
Fetching your mailbox content and steps to execute program:

We need '.mbox' format for our spam folder to run this program. Mbox is a mailbox format that is used by leading email service providers. Its documentation is present in references section of the report. This program will work for any mbox file from any service provider. For this report, I have chosen gmail as it is the most widely used email server. To get the .mbox version of a spam folder, we need to follow below steps for Gmail. For any other server as well, we can get .mbox file easily.

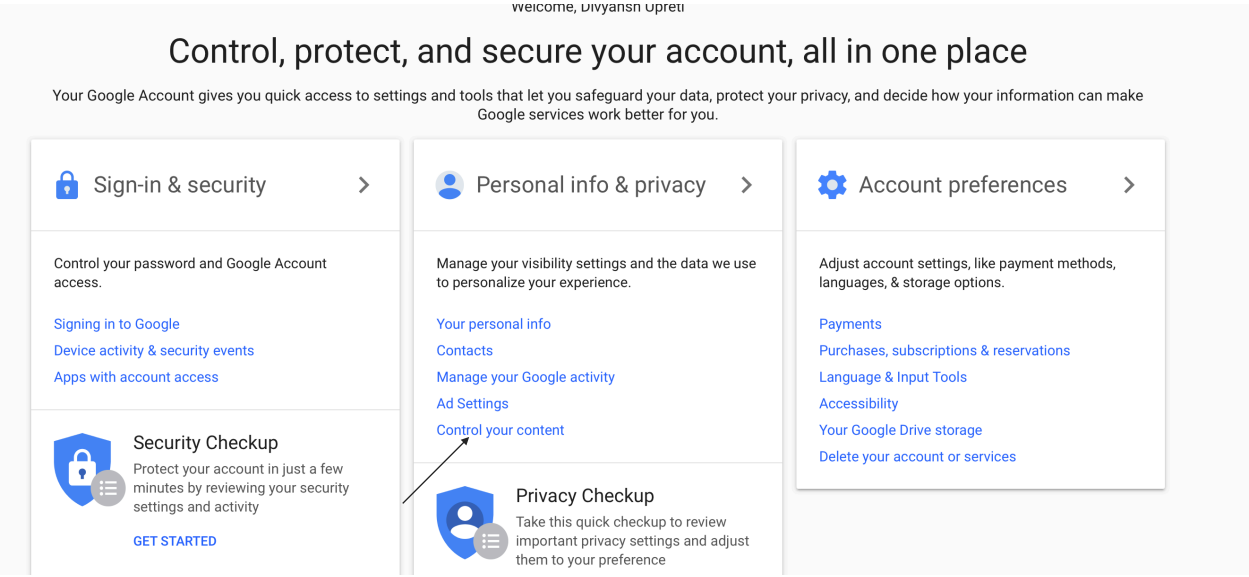
1. Go to google.com



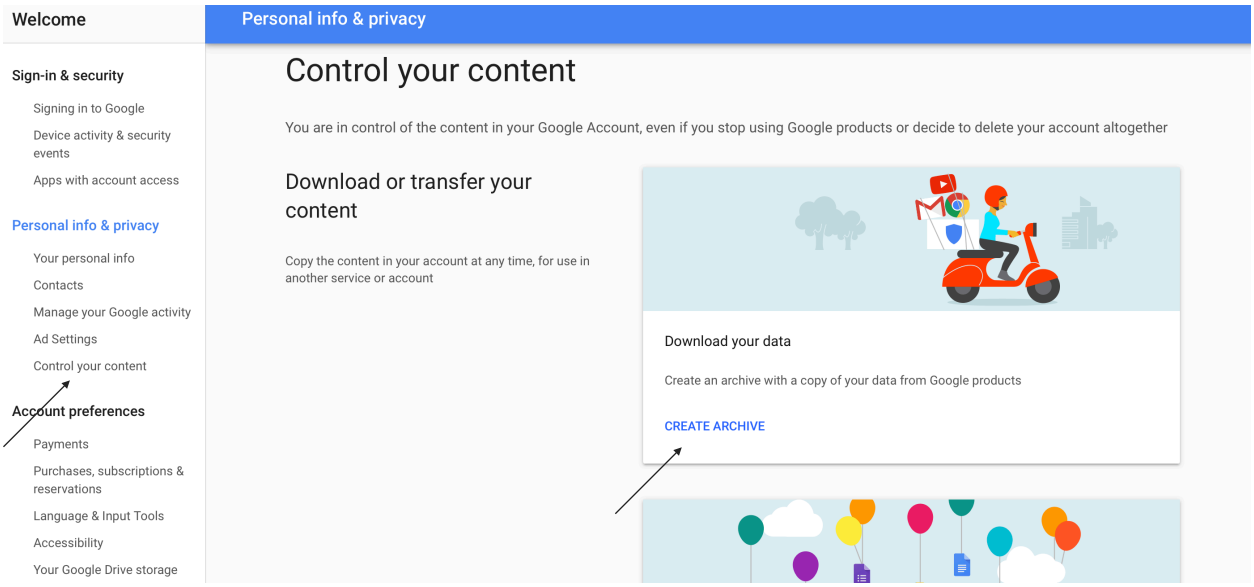
2. Select your google account from the upper right hand corner.
3. In the upper right hand corner, select the menu as shown in figure and click on the first option 'Account'.



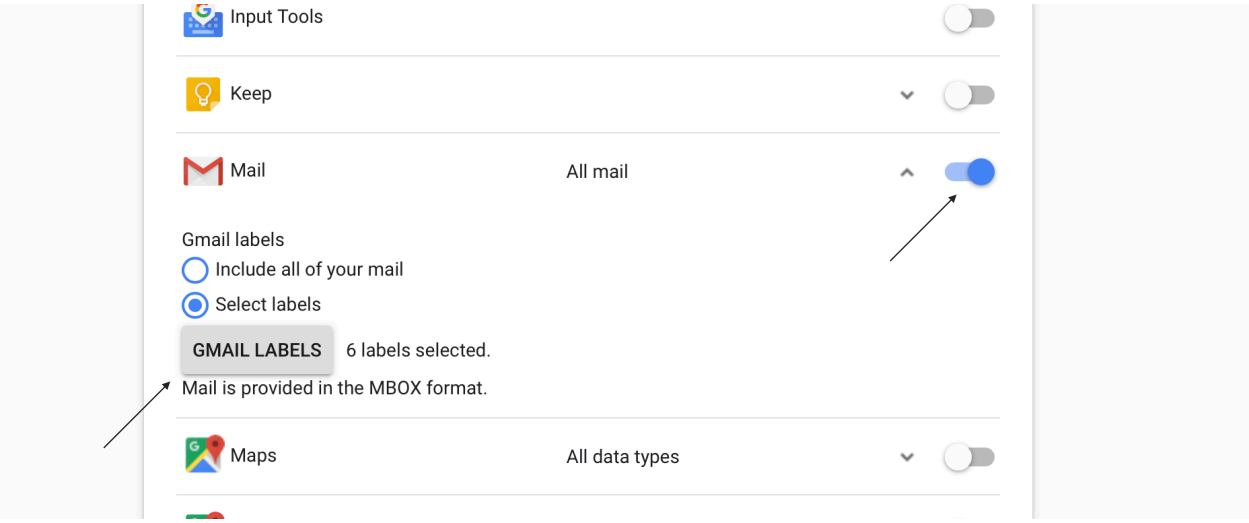
3. Now click on option ‘Contol your content’ as show below.



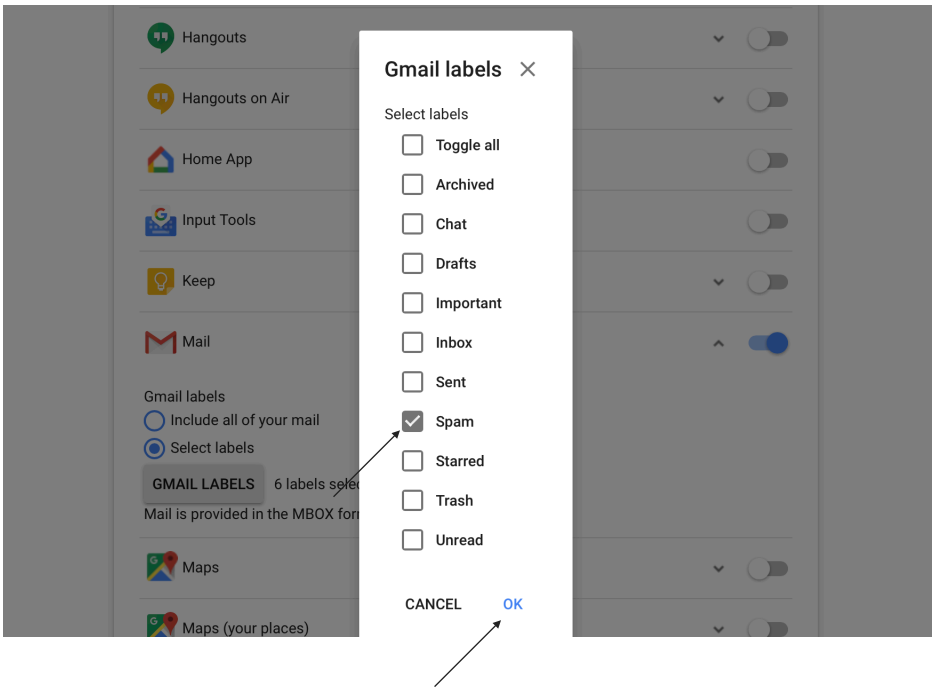
4. With ‘Control your content’ selected in left pane, click on option ‘Create Archive’ in right window.



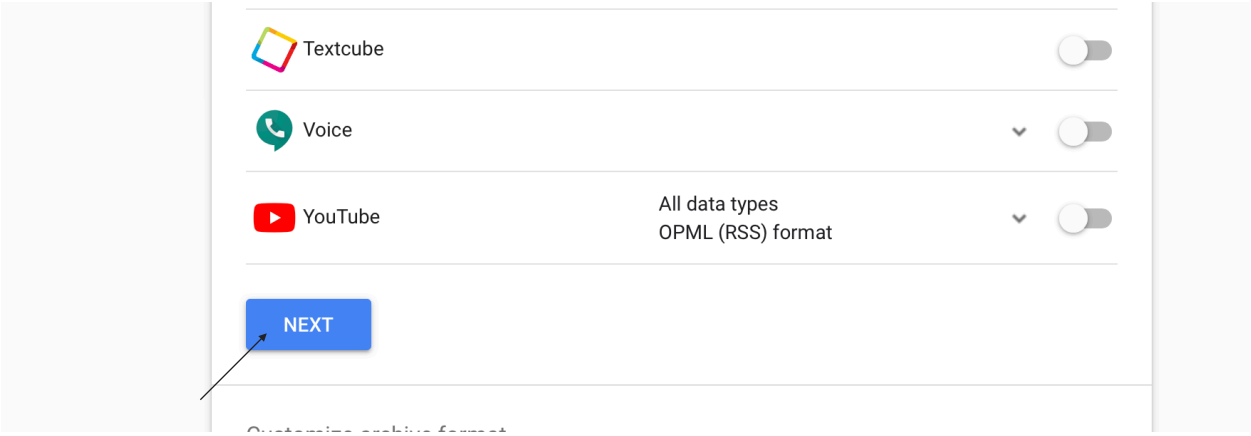
5. With all other options unchecked, select only the mail option and click on option 'Gmail Labels’ as shown below.



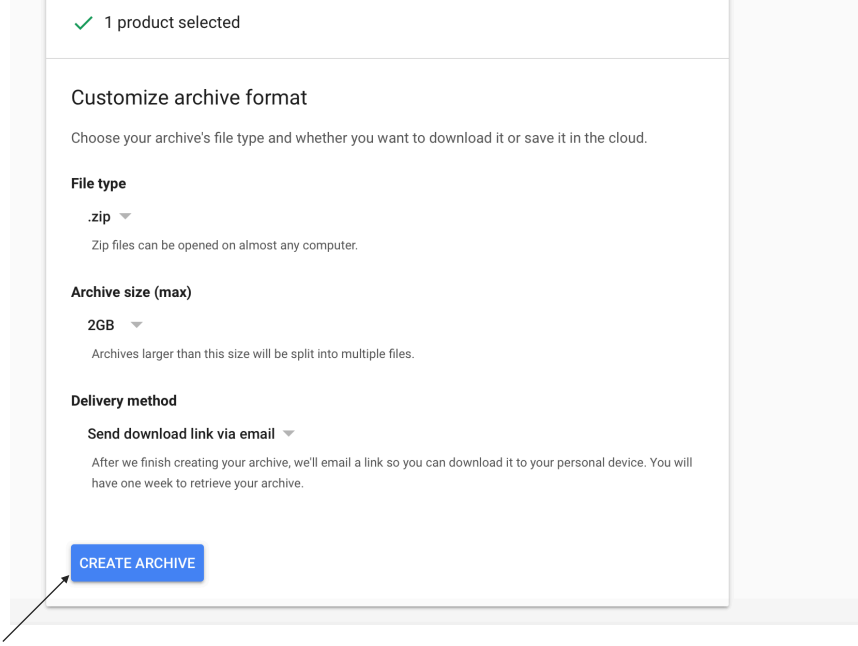
6. With all other options unchecked, select the Spam option as shown below and click on Ok:



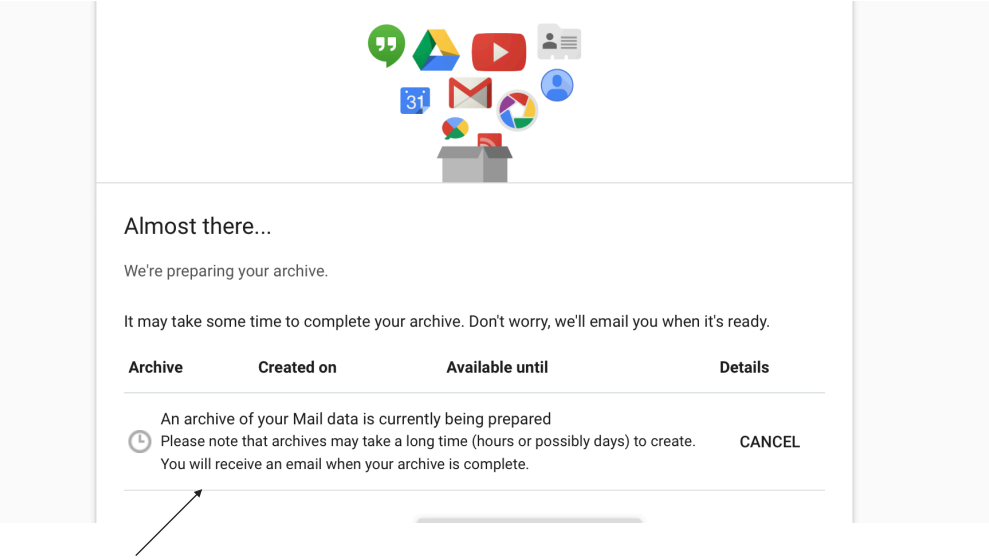
7. Scroll and click on ‘Next’.



8. Click on ‘Create Archive’.



9. Following confirmation window will open. Wait for sometime and the archive with .mbox file will be emailed to you.



10. Once you receive the email, download the .mbox file and put it in the directory of this project along with file 'spamReporter.py'.

11. Run spamReporter.py by the following command:

    sudo python spamReporter.py

12. Required libraries have been mentioned in the README file. Please install them before running this program. This program is developed using Python 3.0. Steps to run the program are also mentioned in the README file.

References:

1. To fetch the word relevant to user's fields/interests, it was important to ignore some other common words used in English. I have ignored the following 100 most common words (information present in the wikipedia page). I also added some words related to email myself.

   Link of 100 most common words: https://en.wikipedia.org/wiki/Most_common_words_in_English

2. Documentation of libraries used:

   https://pythonhosted.org/PyPDF2/

https://pypi.org/project/fpdf/
https://matplotlib.org
http://www.numpy.org
https://pygeoip.readthedocs.io/en/v0.3.2/
https://docs.python.org/3.4/library/re.html
https://docs.python.org/3/library/socket.html
https://docs.python.org/2/library/webbrowser.html
https://docs.python.org/2/library/mailbox.html

3. Parsing the email body :
https://stackoverflow.com/questions/17874360/python-how-to-parse-the-body-from-a-raw-email-given-that-raw-email-does-not

4. Mailbox documentation:
https://en.wikipedia.org/wiki/Mbox