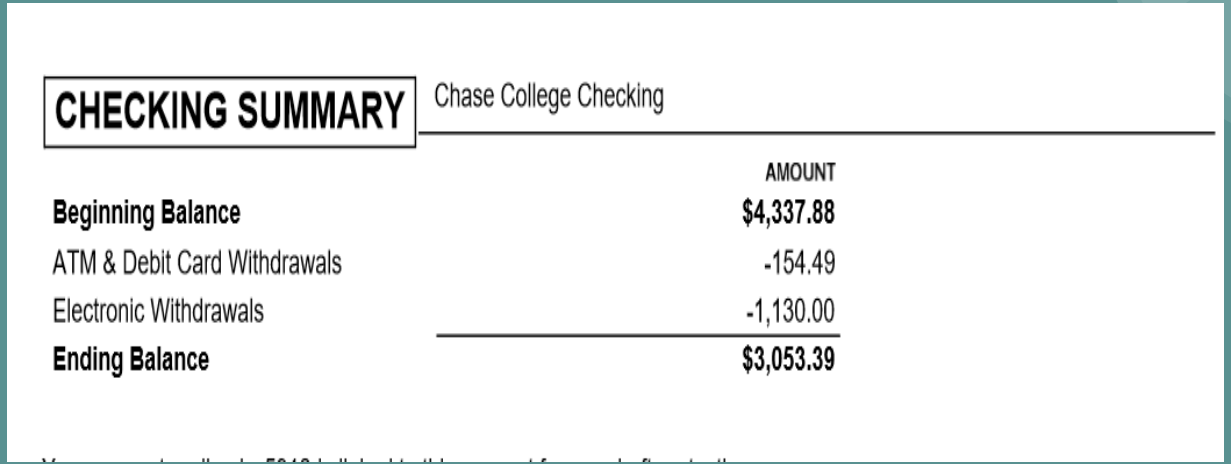
Problem statement

- Scrap bank statements to obtain summary and transaction data.
- Save it in the master CSV.
- Perform data analysis.

Data

Data are bank statements of the **JP Morgan Chase** bank.



CHECKING SUMMARY Chase College Checking

	AMOUNT
Beginning Balance	**$4,337.88**
ATM & Debit Card Withdrawals	-154.49
Electronic Withdrawals	-1,130.00
Ending Balance	**$3,053.39**

CHASE

JPMorgan Chase Bank, N.A.
P O Box 182051
Columbus, OH 43218-2051

[redacted]

[redacted]

CHECKING SUMMARY

Chase College Checking

	AMOUNT
Beginning Balance	**$4,337.88**
ATM & Debit Card Withdrawals	-154.49
Electronic Withdrawals	-1,130.00
Ending Balance	**$3,053.39**

Your account ending in 5912 is linked to this account for overdraft protection.

Cash bonuses paid to this account during 2018 totaled $100.00. This amount will be added to any interest paid during 2018 for tax reporting purposes.

TRANSACTION DETAIL

DATE	DESCRIPTION	AMOUNT	BALANCE
	Beginning Balance		**$4,337.88**
01/22	Quickpay With Zelle Payment To 8312959358 Jpm202538579	-400.00	3,937.88
01/22	Card Purchase With Pin 01/21 Safeway Store 2607 Santa Cruz CA Card 8744	-9.96	3,927.92
01/22	Recurring Card Purchase 01/20 Leetcode.Com Httpsleetcode CA Card 8744	-35.00	3,892.92
01/24	Card Purchase 01/22 USA*Snack Soda Vending Santa Cruz CA Card 8744	-1.00	3,891.92
01/25	Card Purchase 01/24 Ucsc Dining Services #2 Santa Cruz CA Card 8744	-6.25	3,885.67
01/25	Card Purchase 01/24 Ucsc Dining Services #2 Santa Cruz CA Card 8744	-4.95	3,880.72
01/28	Card Purchase With Pin 01/25 Safeway Store 2607 Santa Cruz CA Card 8744	-10.21	3,870.51
01/28	Card Purchase With Pin 01/26 Ucsc Bay Tree Bkstore Santa Cruz CA Card 8744	-1.74	3,868.77
01/28	Quickpay With Zelle Payment To 8312959358 Jpm203798366	-730.00	3,138.77
01/28	Card Purchase 01/28 Coinmach Northern Calif Union City CA Card 8744	-2.75	3,136.02
01/28	Card Purchase 01/28 Coinmach Northern Calif Union City CA Card	-2.75	3,133.27

Information about the Data

- It has **non tabular summary data** which consists of labels and amount.
- In summary, number of labels vary from statement to statement.
- All the statements has different number of pages.
- Transaction data is stored in **semi tabular form.**
- Transaction data starts from different pages in the statements.

Results of using each library

1) PyPDF2 - It returns encoded characters which are difficult to understand. It counts number of pages the PDF has correctly. - 4% approx
2) Tabula : It combines date and description column. It scrapes differently for the pages which has tables from the start and pages which has data and then tables. - 20% approx.
3) PDF miner six - Converts pdf data to array of strings with lot of encoded characters. It combine amount, balance and description as one string without any spaces- 25% approx.

Results of using each library

5) Tikka : It gives result similar has pdf mier six. - 25% approx.

6) PDFQuery: It gives exception as it sees barcode and encoded information in the start. - 0% approx

7) PDF tables: It is a paid library. It scrapes transaction data quite accurately. But not summary data. - 45% approx.

8) Camelot: It completes skips first page and scrapes from the pages which has tables.- 15% approx.

Solution

- Combine libraries
- Use regular expressions
- Lot of string and list manipulations.

Using 3 libraries:
We get Date, description, amount and balance.
We get labels and amount for summary

It is 80% accurate.

All these data is stored in the master CSV:
TransactionsMasterCSV
SummaryMasterCSV


PDF
miner
six


Tabula


PyPDF2

Three winners!!!

1) PDF miner size: Obtain date, summary labels and amount, startpage for the transactions.

1) Tabula: Obtain description, amount and balance.

Use the start page to scrap different for the first page and other pages.

3) PyPDF2 - Obtain page numbers.

Code References

https://stackoverflow.com/questions/39854841/pdfminer-python-3-5/40877143

https://www.zevross.com/blog/2014/04/09/extracting-tabular-data-from-a-pdf-an-example-using-python-and-regular-expressions/

https://towardsdatascience.com/python-for-pdf-ef0fac2808b0

Data scraping tasks

- Try 8 different libraries, to analyze which works efficiently.
- Combine results of three libraries as mentioned in the algorithm to get two master CSVs.

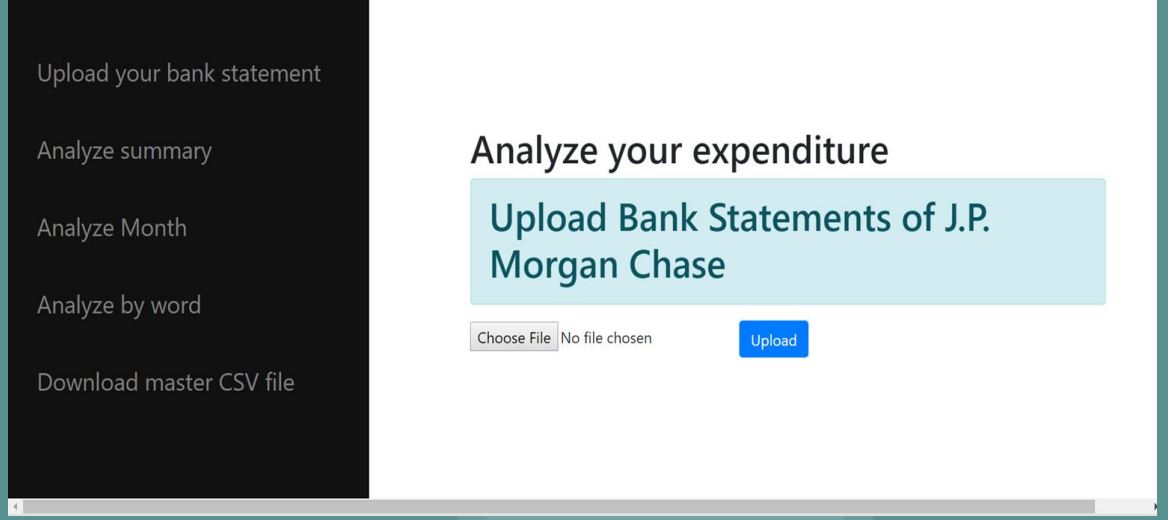
Data cleaning tasks

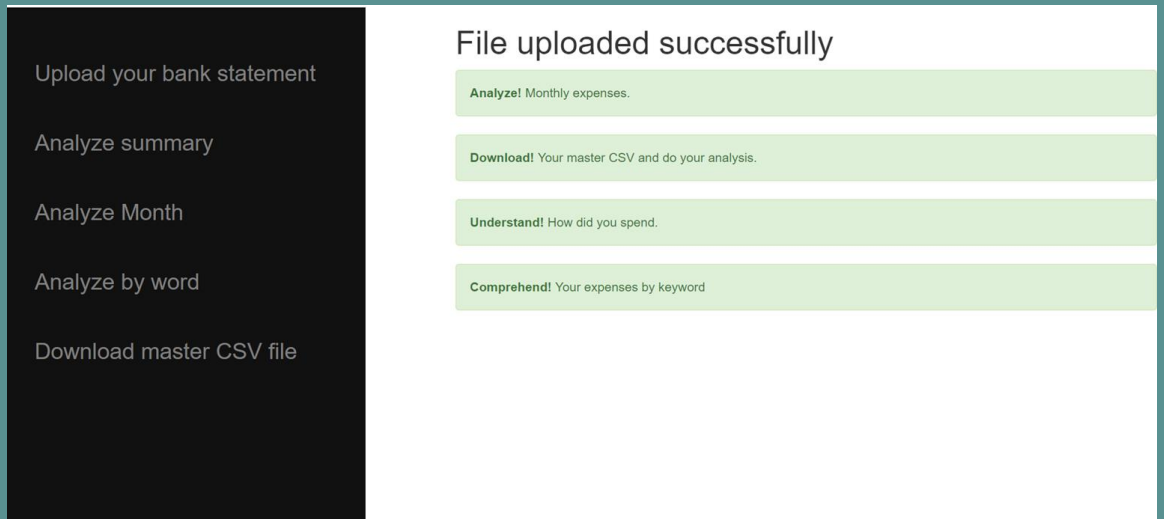
- Clean the data to remove encoded characters.
- Remove nans,new line, tab and space characters.
- Remove $ and commas from the money, to convert them into float.

Analysis and visualization tasks:

- Analyse results of all 8 libraries, to decide which combination can give highest accuracy.
- Use summary csv, to understand the summary by bar graph.
- Use transaction csv, to understand monthly expenditure by line chart.
- Analyze how much you spend on anything by sending keyword.

Built Flask UI to upload





bankstatements.csv

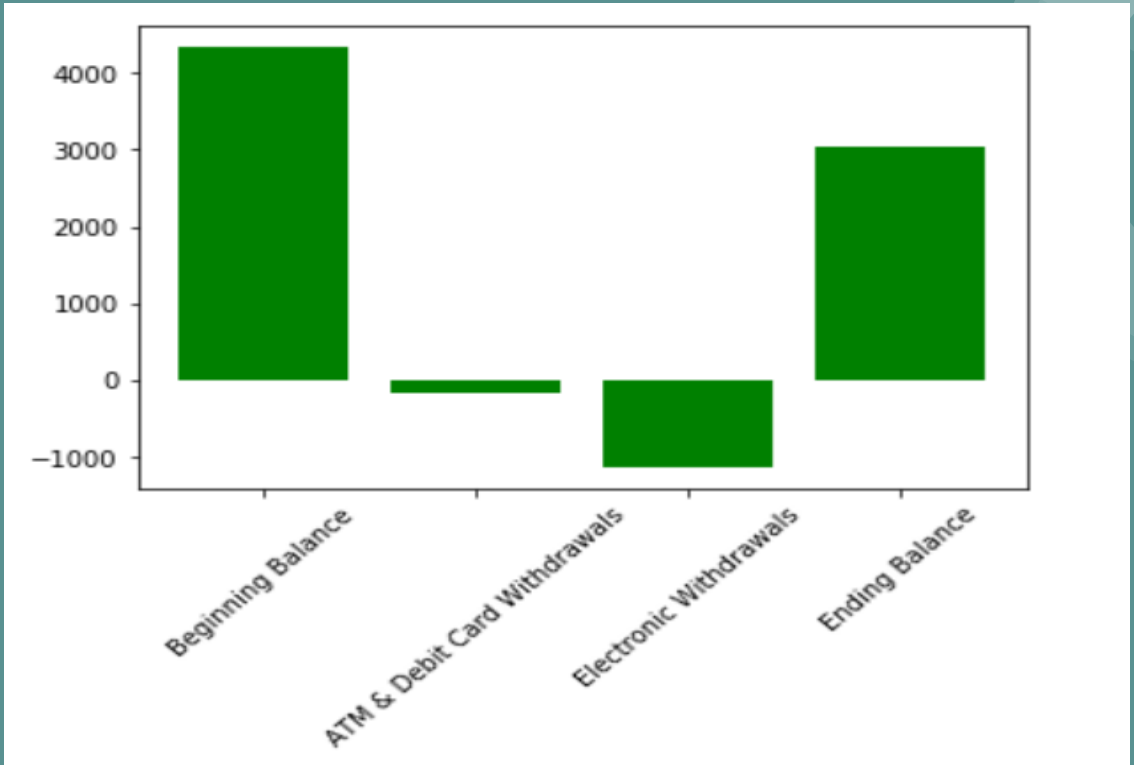
Dates	Months	Description	Amount	Balance
22-Jan	1	Quickpay With Zelle Payment To 8312959358 Jpm202538579	-400	3937.88
22-Jan	1	Card Purchase With Pin 01/21 Safeway Store2607 Santa Cruz CA Card	-9.96	3927.92
22-Jan	1	Recurring Card Purchase 01/20 Leetcode.Com Httpsleetcode CA Card 8744	-35	3892.92
24-Jan	1	Card Purchase01/22 USA*Snack Soda Vending Santa Cruz CA Card	-1	3891.92
25-Jan	1	Card Purchase01/24 Ucsc Dining Services #2 Santa Cruz CA Card	-6.25	3885.67
25-Jan	1	Card Purchase01/24 Ucsc Dining Services #2 Santa Cruz CA Card	-4.95	3880.72
28-Jan	1	Card Purchase With Pin 01/25 Safeway Store2607 Santa Cruz CA Card	-10.21	3870.51
28-Jan	1	Card Purchase With Pin 01/26 Ucsc Bay Tree Bkstore Santa Cruz CA Card	-1.74	3868.77
28-Jan	1	Quickpay With Zelle Payment To 8312959358 Jpm203798366	-730	3138.77
28-Jan	1	Card Purchase01/28 Coinmach Northern Calif Union City CA Card	-2.75	3136.02
28-Jan	1	Card Purchase01/28 Coinmach Northern Calif Union City CA Card	-2.75	3133.27
28-Jan	1	Card Purchase01/28 Coinmach Northern Calif Union City CA Card	-2.5	3130.77
28-Jan	1	Card Purchase01/28 Coinmach Northern Calif Union City CA Card	-2.5	3128.27
4-Feb	2	Card Purchase02/02 Comcast Californ Cs 1 800-266-2278 CA Card	-57.03	3071.24
5-Feb	2	Card Purchase02/04 USA*Snack Soda Vending Santa Cruz CA Card	-1	3070.24
5-Feb	2	Card Purchase02/04 USA*Snack Soda Vending Santa Cruz CA Card	-1	3069.24
11-Feb	2	Card Purchase With Pin 02/10 Safeway Store	-14.85	3054.39
13-Feb	2	Card Purchase02/12 USA*Snack Soda Vending Santa Cruz CA Card	-1	3053.39
15-Feb	2	Quickpay With Zelle Payment To Harshil.Jain-Ves.Ac.IN Jpm208437777	-12	3041.39

Summary csv

Labels	Amount
Month	1
Beginning Balance	4337.88
ATM & Debit Card Withdrawals	-154.49
Electronic Withdrawals	-1130
Ending Balance	3053.39
Labels	Amount
Month	2
Beginning Balance	3053.39
Deposits and Additions	2099.31
ATM & Debit Card Withdrawals	-457.01
Electronic Withdrawals	-742.59
Ending Balance	3953.1
Labels	Amount
Month	4
Beginning Balance	4371.72
Deposits and Additions	2299.31
ATM & Debit Card Withdrawals	-1765.58
Electronic Withdrawals	-4100
Ending Balance	805.45

Summary

Analyze summary



Date wise expenses in a month

Upload your bank statement

Analyze summary

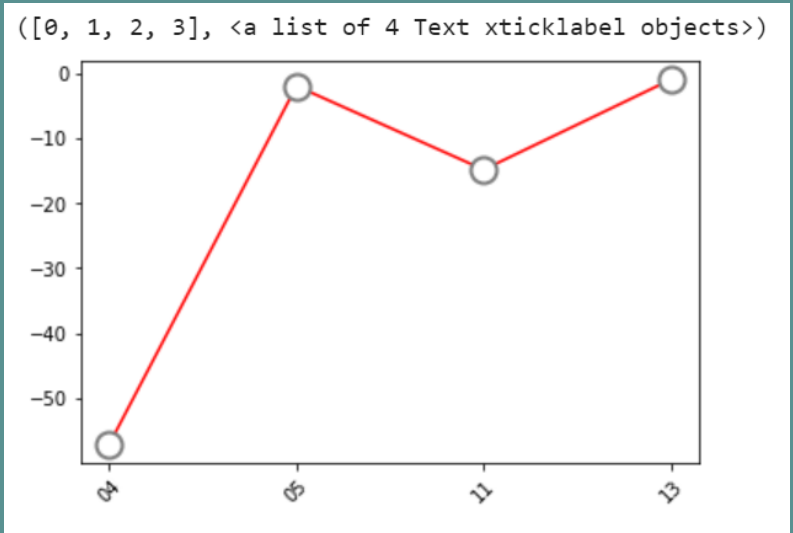
Analyze Month

Analyze by word

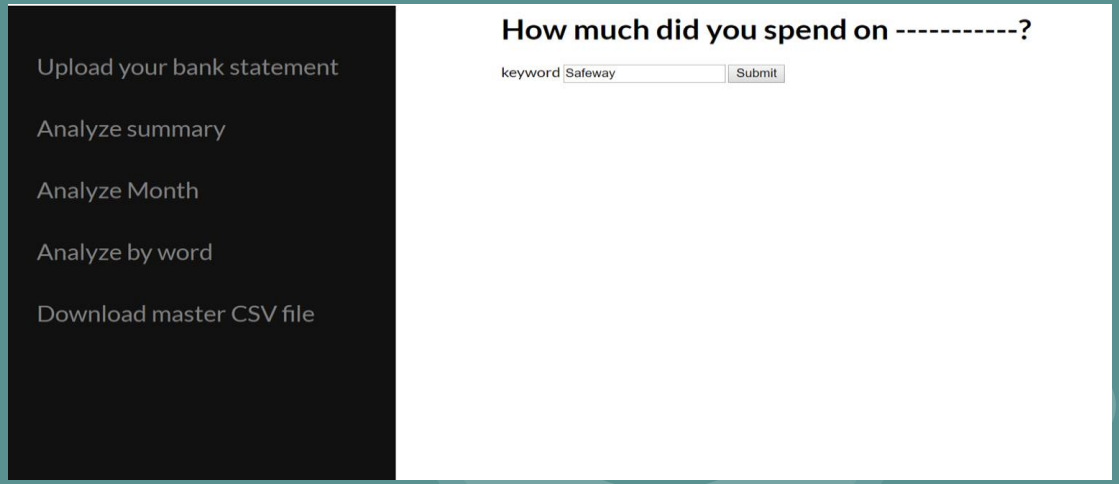
Download master CSV file

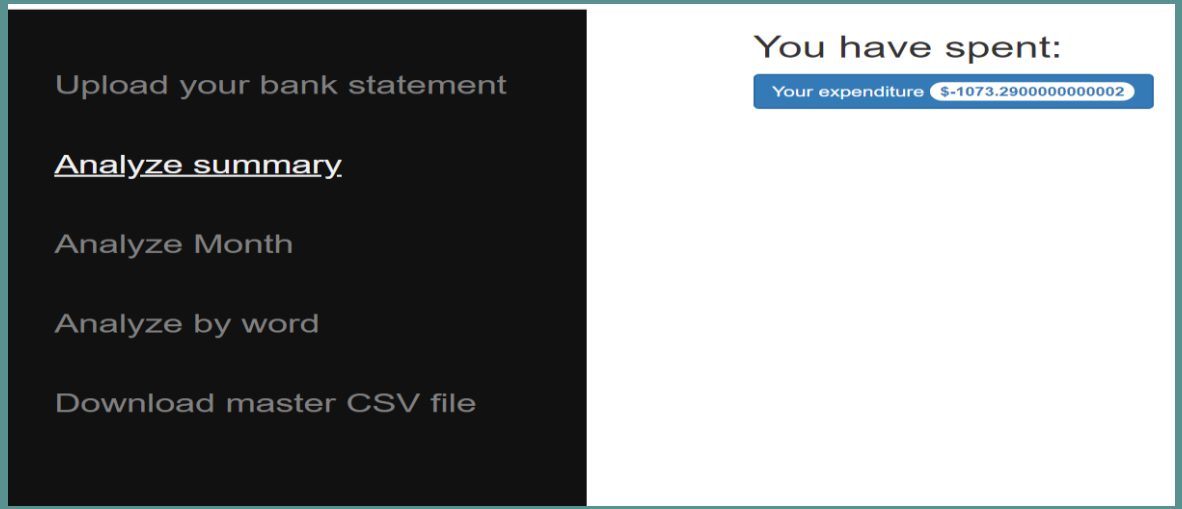
Monthly expense

month 2 Submit



Search your bank statements by keyword





Github repo:

https://github.com/NidhiPankajShah/BankStatementScraping