Text to Data: Make Your Code Malleable, Not Brittle

Martha L Escobar-Molano and David Barrett

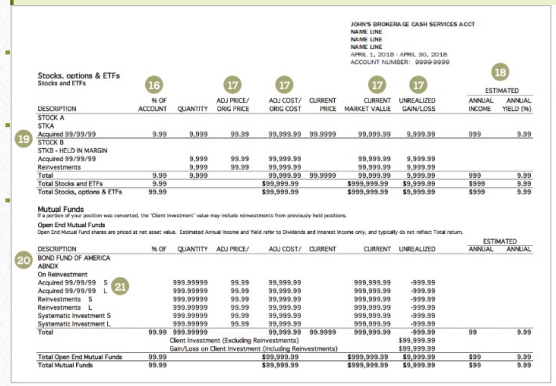
PyData Global 2022

Asgard Systems

Outline

- Problem statement
- Our Approach
- Tagging
- Querying
- Projection
- Jupyter Notebook
- Conclusions

Data Extraction Problem



JOHN'S BROKERAGE CASH SERVICES ACCT
NAME LINE
NAME LINE
NAME LINE
APRIL 1, 2018 - APRIL 30, 2018
ACCOUNT NUMBER: 9999-9999

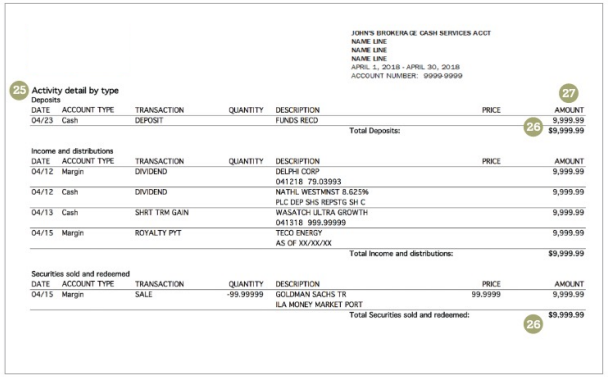
Stocks, options & ETFs
Stocks and ETFs

DESCRIPTION	% OF ACCOUNT	QUANTITY	ADJ PRICE/ ORIG PRICE	ADJ COST/ ORIG COST	CURRENT PRICE	CURRENT MARKET VALUE	UNREALIZED GAIN/LOSS	ESTIMATED ANNUAL INCOME	ESTIMATED ANNUAL YIELD (%)
STOCK A									
STKA									
Acquired 99/99/99	9.99	9,999	99.99	99,999.99	99.9999	99,999.99	9,999.99	999	9.99
STOCK B									
STKB - HELD IN MARGIN									
Acquired 99/99/99		9,999	99.99	99,999.99		99,999.99	9,999.99		
Reinvestments		9,999	99.99	99,999.99		99,999.99	9,999.99		
Total	9.99	9,999		99,999.99	99.9999	99,999.99	9,999.99	999	9.99
Total Stocks and ETFs	9.99			$99,999.99		$999,999.99	$9,999.99	$999	9.99
Total Stocks, options & ETFs	99.99			$99,999.99		$999,999.99	$9,999.99	$999	9.99

Mutual Funds
If a portion of your position was converted, the "Client Investment" value may include reinvestments from previously held positions.

Open End Mutual Funds
Open End Mutual Fund shares are priced at net asset value. Estimated Annual Income and Yield refer to Dividends and Interest Income only, and typically do not reflect Total return.

DESCRIPTION	% OF	QUANTITY	ADJ PRICE/	ADJ COST/	CURRENT	CURRENT	UNREALIZED	ESTIMATED ANNUAL	ESTIMATED ANNUAL
BOND FUND OF AMERICA									
ABNDX									
On Reinvestment									
Acquired 99/99/99 S		999.99999	99.99	99,999.99		999,999.99	-999.99		
Acquired 99/99/99 L		999.99999	99.99	99,999.99		999,999.99	-999.99		
Reinvestments S		999.99999	99.99	99,999.99		999,999.99	-999.99		
Reinvestments L		999.99999	99.99	99,999.99		999,999.99	-999.99		
Systematic Investment S		999.99999	99.99	99,999.99		999,999.99	-999.99		
Systematic Investment L		999.99999	99.99	99,999.99		999,999.99	-999.99		
Total	99.99	999.99999		99,999.99	99.9999	999,999.99	-999.99	99	9.99
		Client Investment (Excluding Reinvestments)					$99,999.99		
		Gain/Loss on Client Investment (Including Reinvestments)					$99,999.99		
Total Open End Mutual Funds	99.99			$99,999.99		$999,999.99	$9,999.99	$99	9.99
Total Mutual Funds	99.99			$99,999.99		$999,999.99	$9,999.99	$99	9.99



JOHN'S BROKERAGE CASH SERVICES ACCT
NAME LINE
NAME LINE
NAME LINE
APRIL 1, 2018 - APRIL 30, 2018
ACCOUNT NUMBER: 9999-9999

Activity detail by type

Deposits

DATE	ACCOUNT TYPE	TRANSACTION	QUANTITY	DESCRIPTION	PRICE	AMOUNT
04/23	Cash	DEPOSIT		FUNDS RECD		9,999.99
				Total Deposits:		$9,999.99

Income and distributions

DATE	ACCOUNT TYPE	TRANSACTION	QUANTITY	DESCRIPTION	PRICE	AMOUNT
04/12	Margin	DIVIDEND		DELPHI CORP		9,999.99
				041218 79.03993		
04/12	Cash	DIVIDEND		NATHL WESTMNST 8.625%		9,999.99
				PLC DEP SHS REPSTG SH C		
04/13	Cash	SHRT TRM GAIN		WASATCH ULTRA GROWTH		9,999.99
				041318 999.99999		
04/15	Margin	ROYALTY PYT		TECO ENERGY		9,999.99
				AS OF XX/XX/XX		
				Total Income and distributions:		$9,999.99

Securities sold and redeemed

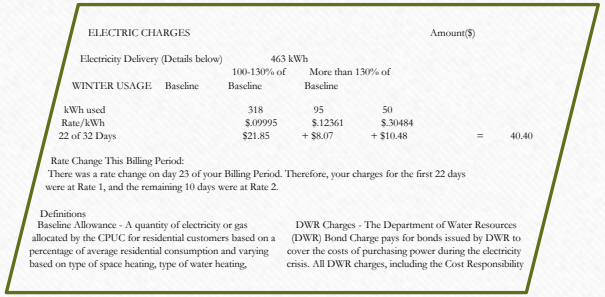
DATE	ACCOUNT TYPE	TRANSACTION	QUANTITY	DESCRIPTION	PRICE	AMOUNT
04/15	Margin	SALE	-99.99999	GOLDMAN SACHS TR	99.9999	9,999.99
				ILA MONEY MARKET PORT		
				Total Securities sold and redeemed:		$9,999.99

Ingest


```
{
   "_id" : ObjectId("5b8ae97f64908e0afcab0fc6"),
   "name" : "JOHN",
   "interval" : [
      "April 1, 2018",
      "April 30, 2018"
   ],
   "holdings" : [
      {
         "account number" : "9999 9999",
         "balance" : [
            {
               "ticker" : "STKA",
               "quantity" : 9,999,
               "current value" : 99,999.99
            },
            {
               "ticker" : "STKB",
               "quantity" : 9,999,
               "current value" : 99,999.99
            },
            {
               "ticker" : "ABNDX",
               "quantity" : 999.99999,
               "current value" : 999,999.99
            }
         ],
         "transactions" : [
            {
               "date" : "04/23",
               "account type": "Cash",
               "transaction": "DEPOSIT",
               "description" : "FUNDS RECD",
               "amount" : 9,999.99
            },
            ...
         ]
      }
   ]
}
```

Approach

TEXT



ELECTRIC CHARGES — Amount($)

Electricity Delivery (Details below) — 463 kWh

WINTER USAGE	Baseline	100-130% of Baseline	More than 130% of Baseline		
kWh used		318	95	50	
Rate/kWh		$.09995	$.12361	$.30484	
22 of 32 Days		$21.85	+ $8.07	+ $10.48	= 40.40

Rate Change This Billing Period:
There was a rate change on day 23 of your Billing Period. Therefore, your charges for the first 22 days were at Rate 1, and the remaining 10 days were at Rate 2.

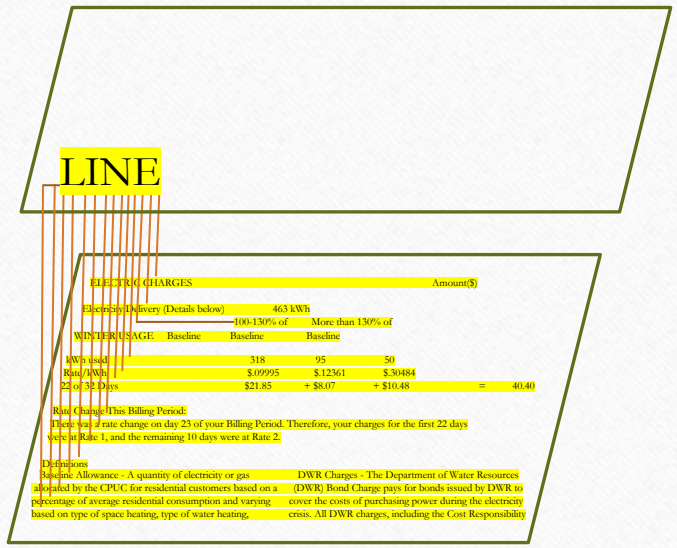
Definitions
Baseline Allowance - A quantity of electricity or gas allocated by the CPUC for residential customers based on a percentage of average residential consumption and varying based on type of space heating, type of water heating,

DWR Charges - The Department of Water Resources (DWR) Bond Charge pays for bonds issued by DWR to cover the costs of purchasing power during the electricity crisis. All DWR charges, including the Cost Responsibility

Approach - Tagging

TAGS

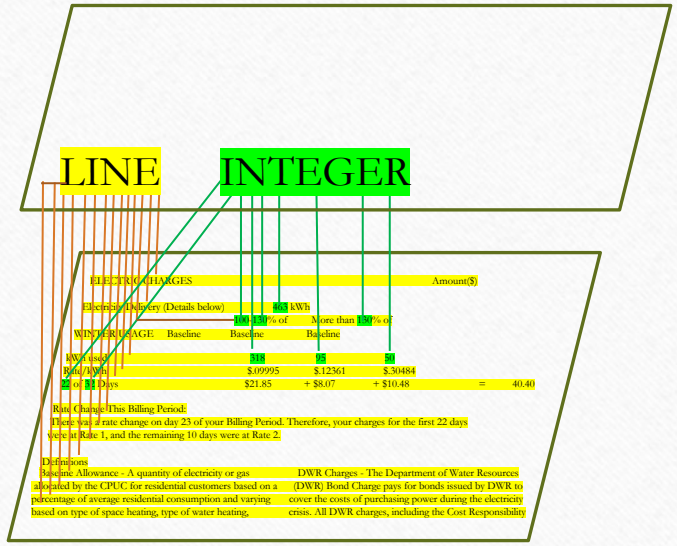
TEXT



Approach - Tagging

TAGS

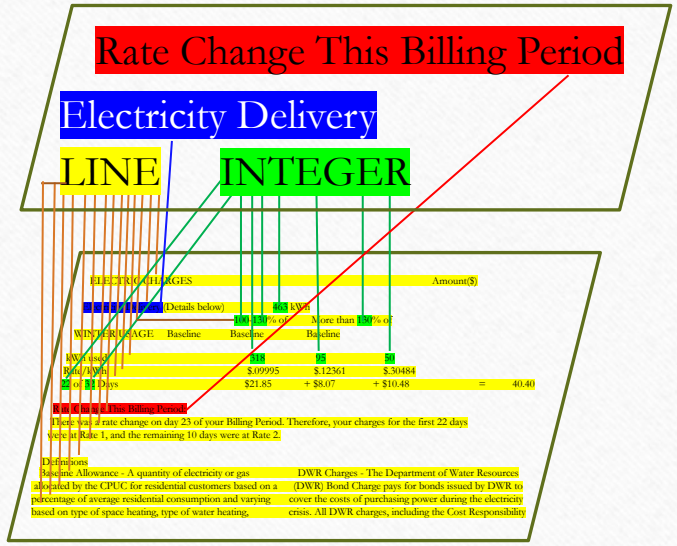
TEXT



Approach - Tagging

TAGS

TEXT



Approach - Querying

QUERY

Select text between LINE starting with 'Electricity Delivery', and string 'Rate Change This Billing Period'

Result

TAGS

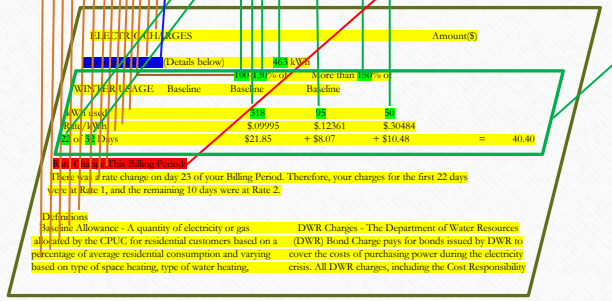
Rate Change This Billing Period

Electricity Delivery

LINE

INTEGER

TEXT



Tagging

- Associate a tag with a list of string locations that match a regular expression.
  - INTEGER is associated with strings in text that match "[0-9]+"
- A string location is defined as a triple: ***o***, ***s***, ***e*** where ***o*** is an offset from beginning of the text, and ***s, e*** are the start and end locations of the string with respect to the offset
- A tag could be anonymous by using the text to be matched in place of a tag
  - 'Electricity Delivery' can be used instead of a tag to denote locations where string 'Electricity Delivery' occurs

Querying

- Once text has been tagged, we can query the tags to retrieve text in the document
- Queries are performed in three steps:
  1. Select set of location tuples
  2. Apply function to each tuple of step 1
  3. Apply function to result of step 2
- The result of step 3 is a list of locations that can be mapped into text in the document

Step 1 – Select Tuples of locations

- To query the document, we first select tuples of locations as follows:

  $\{(loc_1, loc_2, \ldots, loc_n) \mid loc_1 \in tag_1 \wedge \cdots \wedge loc_n \in \text{tag}_n \wedge \text{predicate}(loc_1, \ldots, locn)\}$

  Where predicate consists of conjunctions and disjunctions of relations on locations.

- Relations can be built-in or **user defined**. Built-in relations include Allen's

before	meets	during	finishes	equal	overlaps	starts
XXX YYY	XXXYYY	XXX YYYYYY	XXX YYYYY	XXX YYY	XXX YYY	XXX YYYYY

Step 1 – Example

Select text between LINE starting with 'Electricity Delivery', and string 'Rate Change This Billing Period'

- Identify three locations where the first is tagged LINE, the second and third are occurrences of 'Electricity Delivery' and 'Rate Change This Billing Period', respectively.
  - Function literal(x) returns the set of locations in the text that matches x.
- Constrain these locations so that second location starts the first one, and the first location appears before the third one

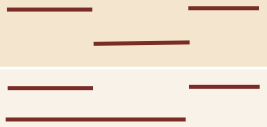
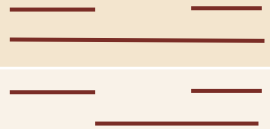
$$\{(loc_1, loc_3) \mid \exists loc_2 \colon loc_1 \in \text{LINE} \wedge loc_2 \in \text{literal}(\text{'Electricity Delivery'}) \wedge loc_3 \in \text{literal}(\text{'Rate Change This Billing Period'}) \wedge \text{starts}(loc_2, loc_1) \wedge \text{before}(loc_1, loc_3)\}$$

- It returns a pair: location of line with blue text and location of red text

Step 2 – Apply function to Tuples

Once location tuples are identified, a function might be applied to each tuple to get desired data. Typically, this function returns a single location:

C: $(\text{loc}_1, \text{loc}_2, \ldots, \text{loc}_n) \rightarrow \text{loc}$

Function	Input/Output	Function	Input/Output
in_between		closed_span	
open_right_span		open_left_span	

To get text between selected LINE and 'Rate Change This Billing Period', we apply in_between to each resulting tuple from step 1.

{ in_between(loc_1, loc_3) | $\exists loc_2$: $loc_1 \in$ 'LINE' $\wedge$ $loc_2 \in$ literal('electricity delivery') $\wedge$ $loc_3 \in$ literal('rate change this billing period') $\wedge$ starts(loc_2, loc_1) $\wedge$ before(loc_1, loc_3) }

Step 3 – Apply function to Result

- Suppose we have the following text:

  ….
  electricity delivery
  ….
  ….
  Rate change this billing period
  ….
  Rate change this billing period

- Result of in-between in previous slide includes 2 locations:
  - of highlighted text (yellow and green)
  - of text highlighted in yellow only
- To eliminate the extra location (the first one), an function needs to be applied to result of step 2

- We need to eliminate extra locations by applying a function to result of Step 2:
  - shortest($\{loc_1, \ldots, loc_n\}$) = $\{loc_i \mid 1 \leq i \leq n \wedge \nexists j\ (1 \leq j \leq n \wedge loc_j \subseteq loc_i)\}$

shortest({ in_between(loc_1, loc_3) | $\exists loc_2$: $loc_1 \in$ 'LINE' $\wedge$ $loc_2 \in$ literal('electricity delivery') $\wedge$ $loc_3 \in$ literal('rate change this billing period') $\wedge$ starts(loc_2, loc_1) $\wedge$ before(loc_1, loc_3) })

Projection

- Locations of city names and state acronyms would not overlap if defined with respect to the beginning of the text

' city state\nBoston ma\nLos Angeles ca\nSan Diego ca\n

3 6 13 17 19 24 34 35 37 47 52 53 55 63 70 71

- They will overlap if locations are defined with respect to beginning of line

' city state\nBoston ma\nLos Angeles ca\nSan Diego ca\n'

3 6 13 17 0 5 15 16 0 10 15 16 0 8 15 16

city	state
Boston	ma
Los Angeles	ca
San Diego	ca

Jupyter Notebook

Asgard Systems