Textract on an invoice:

With standard raw text reading for this file: <https://drive.google.com/file/d/1g1XLrNR0OoMXBoTNhUV5MjeM1f-fxu89/view?usp=sharing>

["CUSTOMER'S ORIGINAL INVOICE CONFIDENTIAL PROPERTY OF SYSCO", 'DKLV DATE', 'CUSTOMER', 'INVOICE NUMBER', 'PAGE', 'MarginEdge Test', 'Sysco', '7/14/20', '8315 Lee Highway. Suite 50', 'TROCK STOP', 'Phone: 703-682-0999 x2', 'Fairfax, Virginia 22031', '4', '1', '84106-3287', 'At the heart of', 'SYSCO INTERMOUNTAIN', '/005', 'food and service', '9494 SOUTH PROSPERITY ROAD', 'ROUTE', 'PURCHASE ORDER', 'MarginEdge Test', 'C', 'WEST JORDAN, UT 84081-0190', '2578', '8315 Lee Highway, Suite 50', '(801) 563-6255 SHIPPER# RS004', 'TERMS -PAST DUE BALANCES ARE SUBJECT TO SERVICE CHARGE', '1-15 DUE 25TH/16-31 DUE 10TH', 'Fairfax, Virginia 22031', 'MANIFEST# 152666 NORMAL DELIVERY', 'MARK WATE', 'UT', '84106 -3287', 'MA: S1C01 LELAND GREG ATKIN', 'QTY', 'PACK', 'SIZE', 'DRIVER: COVERT', 'ITEM DESCRIPTION', 'ITEM', 'UNIT', '-', 'EXTENDED', 'INVOICE ADJUSTMENTS', 'CODE', 'PRICE', 'AMOUNT', 'PRICE', 'CODE', 'QTY', 'ACC', 'DAIRY PRODUCTS', 'c', '1 CS', '10', 'O#AVGRBRLIMP CHEESE CHEDDAR SHARP PRIN SYS2822312', '2822312', '4.316', '10.640', '45.92', 'T/WT=', '10.640', 'C', '1S', "ONLY'S LB", 'BBRLCLS CHEESE SWISS/AMER 120 SLI', '14716', '5148453', '17.79', '17.79', 'GROUP TOTAL\*\*\*\*', '63.71', 'ACTIVE', 'POULTRY', 'C', '1', 'CS', '410 LB', 'SYS CLS CHICKEN CVP WING 142JT JMB RND 52890', '6344790', '87.62', '87.62', 'CLAIM', 'GROUP TOTAL', '87.62', 'CANNED & DRY', 'OF', ':', 'D', '1SCS', '123 LB', 'MORTON SALT KOSHER', '1702', '1995125', '33.33', '33.33', '9', 'D', '1S', 'ONLY1 8 oz', 'IMP/MCC SEASONING CAJUN', '974235', '5228424', '19.21', '19.21', '$', 'GROUP TOTAL', '52.54', 'PRODUCE', '.', '$', 'C', '1', 'CS', '11 LB', 'IMPFRSH DILL BABY FRESH HERB', '2005148', '15.25', '15.25', 'ME', 'C', '1', 'CS', '120 LB', 'PACKER CUCUMBER PICKLING FRESH', '2034023', '37.28', '37.28', '2', 'CS', '15 LB', 'PACKER CARROT BABY PLD TRI COLOR', '7680291', '32.50', '65.00', 'GROUP TOTAL', '117.53', 'MISC CHARGES', 'CHGS FOR FUEL SURCHARGE', '3.50', '.', 'AND', 'ORDER SUMMARY', ': 1277265', 'SECURITY', 'CASES', 'SPLIT TOT PCS', 'CUBE', 'GROSS WT.', 'OPEN:10:30 AM', 'CLOSE:', '6:00 PM', 'REMIT TO', 'SYSCO INTERMOUNTAIN, INC', 'SUB', '7', '2', 'g', '132', 'P.O. BOX 190', '4.9', 'TOTAL', 'WEST JORDAN, UTAH 8408', '324.90', 'TAX', '7', '2', '9', '4.9', '132', 'CUST INVOICE EVED or ALL STAMP', 'NO. PCS', 'TOTAL', "DRIVER'S &", 'NO PCS', 'SIGN', 'DELVD', 'SIGN', 'REC.', 'INVOICE', 'X', 'TOTAL', '324.90', 'PROVISION THE PERISHABLE THE FERISHABLE AGRICULTURAL AGRIC COMMODITIES LISTED ON THIS INVOICE ARE SUBJECT THE SELLER TO THE or STATUTORY THIS COMMODITY TRUST PAYABLE ON OR BEFORE', 'ALL LTURAL COMMODITIES ACT 1930 FROM THESE COMMODITIES', '7/25/20', 'LAST PAGE', 'PROCEEDS FROM MAIL THESE INVENTORIES OTHER PRODUCTS DERIVED RECEIVED FURTHER YOU AGREE WITH', 'ARTSING OUT or YOUR RECEIPT OF COMMODITIES UNTIL FULL GIVING UP YOUR RIGHT TO SERVE IN ANY', 'PAYMENT', 'IS', 'CAPACITY, OR TO PARTICIPATE AS A MEMBER OF THESE CLASS PRODUCT OF CLAIMANTS /SERVICES IN YOU ANY ARE LAWSUIT INVOLVING ANY SUCE DISPUTE']

Easily parsable information is needed when considering that all invoices can potentially have different designs and formats:

* Date - regular expression
* Time - Regular expression
* Name - dictionary of all possible places that they shop and use “in” to check every word
* Total - “in” and then check the next item in it.

Difficult to parse every item due to the formatting differences between all invoices, code would have to be hardcoded to every location which will take a long time. Additionally, we severely lack the sample data to do this coding.

* Possible way to do this is to textract it into a tabular format, check the titles of the column headers and then see if the they correspond to hardcoded names like “item”, “cost”, “weight”, “count”, “qty”, “price”, etc

If so then parse that column, otherwise, ignore the column.

This would also require multiple pieces of sample data as different company’s invoices will have different column names and/or abbreviations.

The invoice in this is not strictly tabular either as there are subheaders for the example above:

“Dairy Products”, “Produce”, etc

It also has “group totals” which is not popular for invoices.