NAME :- MOHAN SAI KOTHAPALLI

EMP ID :- 2112951

COHORT CODE :- GN22CDBDS001

ASSIGNMENT ON REGEX

Design a python program to accept a file name through command line arguments.

Parse this file to perform the following:

1. Print all currencies in text, Accepted- $, ₹, £
2. Print all date times in the text- dd/mm/yyyy, dd/mm/yy, mm/dd/yyyy, mm/dd/yy
3. Print all cardinilities and orders- 4th, fifth, sixth, 1st, 2nd, nineteenth, fifth
4. Print all 4 letter words that begin with vowels

STEP 0 :- IMPORTING LIBRARIES

In [ ]:
```
import re
import sys
```

STEP 1 :- Opening DATA FILE passed as an CMD LINE ARGUMENT

In [ ]:
```
f=open(sys.argv[1])
```

In [ ]:
```
with open(sys.argv[1],'r',encoding='utf-8') as f:
    data=f.read()
data[:1000]
```

STEP 2:- FINDING CURRENCY'S INCLUDING THE AMOUNT IN TEXT

In [ ]:
```
x=re.findall(r"(\d*?\.?\d+ ?[$₹£])",data)
x1=re.findall(r"([$₹£] ?\d*\.?\d+)",data)
print(x+x1)
```

FINDING ONLY THE SYMBOLS OF CURRENCY IN TEXT

In [ ]:
```
curr=re.findall("([$₹£])",data)
print("Total Number Of Currency Symbols In the TEXT DATA are : ",len(curr))
print(f"Types Of Currency Symbols In the TEXT DATA are : ",len(set(curr))," ",set(curr))
print(curr)
```

STEP 3 :-PRINTING ALL THE FORMATES OF DATES IN THE TEXT

In [ ]:
```
dates=re.findall(r"((0[1-9]|1[0-2])/(0[1-9]|[12][0-9]|3[01])/(\d{4})\b)",data)
print("The Number of dates in the format of 'mm/dd/yyyy' are : ",len(dates))
for i in range(len(dates)):
    print(dates[i][0],end='  ')
print()
dates=re.findall(r"((0[1-9]|[12][0-9]|3[01])/(0[1-9]|1[0-2])/(\d{4})\b)",data)
print("The Number of dates in the format of 'dd/mm/yyyy' are : ",len(dates))
for i in range(len(dates)):
    print(dates[i][0],end='  ')
print()
dates=re.findall(r"((0[1-9]|[12][0-9]|3[01])/(0[1-9]|1[0-2])/(\d{2})\b)",data)
print("The Number of dates in the format of 'dd/mm/yy' are : ",len(dates))
for i in range(len(dates)):
    print(dates[i][0],end='  ')
print()
dates=re.findall(r"((0[1-9]|1[0-2])/(0[1-9]|[12][0-9]|3[01])/(\d{2})\b)",data)
print("The Number of dates in the format of 'mm/dd/yy' are : ",len(dates))
for i in range(len(dates)):
    print(dates[i][0],end='  ')
print()
```

STEP 4 :- PRINTING ALL CARDINILITIES AND ORDERS FROM THE TEXT

In [ ]:
```
order=[]
x=re.findall(r"((first|second|third|sixth)|(thir[a-z]+(th|st|nd))|(fou[a-z]+(th|rd|st|nd))|(fif[a-z]+(th|st|nd|rd))|(fi[a-z]+th)|(six[a-z]+(th|st|nd|rd))|(sev[a-z]+(th|st|nd|rd))|(eig[a-z]+(th|st|nd|rd))|(nine[a-z]+(th|st|nd|rd))|(ten[a-z]+?(th|st|nd|rd))|(ele[a-z]+th)|(twe[a-z]+(th|st|nd|rd))|(hun[a-z]+(th|st|rd)))",data)
for i in range(len(x)):
    if(x[i][0] not in order):
        order.append(x[i][0])
#print(order)
orders=[]
x1=re.findall(r"([0-9]+(th|st|nd|rd))",data)
for i in range(len(x1)):
    if(x1[i][0] not in orders):
        orders.append(x1[i][0])
#print(orders)
if order or orders:
    print("The cardinilites and orders present in the data are :\n",order+orders)
else:
    print("No CARDINILITIES AND ORDERS FOUND IN THE TEXT")
```

STEP 5 :- PRINTING ALL 4 LETTER WORDS STARTING WITH VOWELS FROM THE TEXT

In [ ]:
```
four=re.findall(r"(\b(a|e|i|o|u|A|E|I|O|U)[a-zA-Z]{3}\b)",data)
fours=[]
if four:
    for i in range(len(four)):
        fours.append(four[i][0])
    print(fours)
    print("Tolal number of words without repetition are : ",len(set(fours)),"\n",set(fours))
else:
    print("NO 4 LETTER WORDS STARTING WITH VOWELS FROM THE TEXT")
```