DISCOVERING HIGH CHOLESTEROL IN FOOD DATA SET

SRIVARSHA BOLLU
810866077
DEPARTMENT OF COMPUTER SCIENCE
KENT STATE UNIVERSITY- KENT -OHIO
sbollu@kent.edu

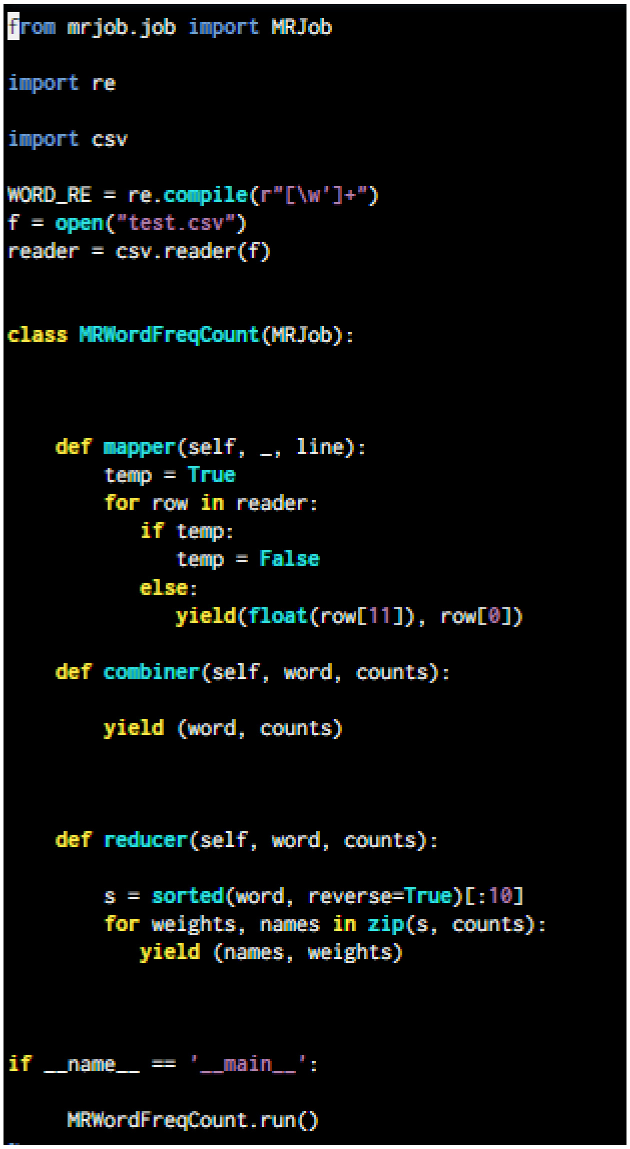
Abstract: Nutrients are the components of food which help to service, we have many type of foods which have high nutrition in it, but many of us doesn't know which food is good or dangerous. This paper will analyze a large data set of nutrients which are having highest fat value. This paper is mainly about analyzing a large data set using Hadoop, the data set is about the list of nutrients that are detailed by the weight, fat value, fat content etc. This paper states about the highest fat value content in the listed food table.

Introduction: The main reason for analyzing this data is to spread the knowledge of harmful food that contains high fat value which leads to many problems. Due to overconsumption of high fat value food contents this may damage our health. So to analyze the list of foods we consume in our daily life we have taken the nutrition's data set and analyze the highest fat content in the listed food.

Methods: To analyze the data set I used Hadoop, MRJOB, python and data set.

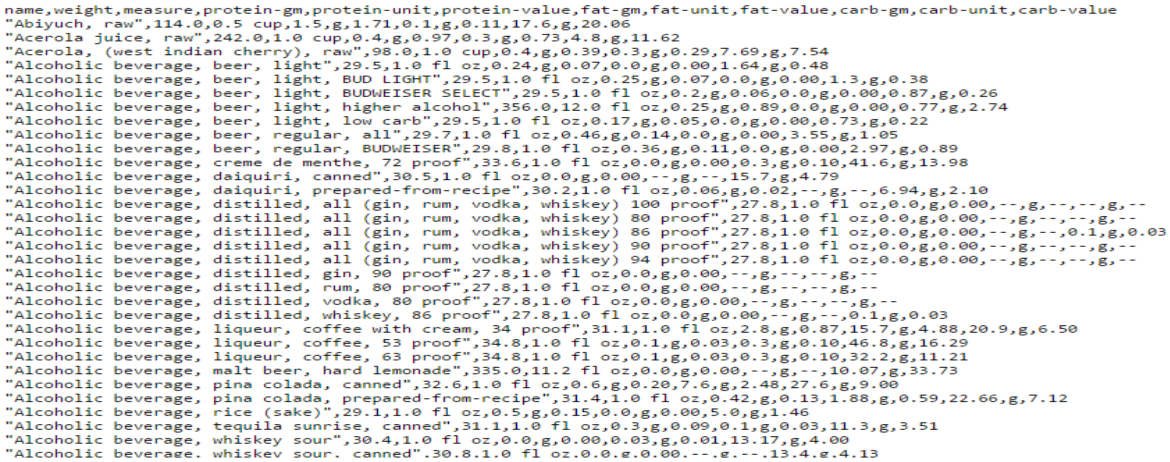
- Hadoop: In Hadoop is the basic requirement to analyze the large data set, here in nutrients data set we implemented on the kent cluster which has Hadoop set up in it.
- Python: python language is used to implement the MRJOB.
- MRJOB: MRJOB does the same thing as MapReduce, it has three main operations in it.
    1. Mapper ()
    2. Combiner ()
    3. Reducer ()
- Dataset: The large data set used is nutrients data set.

MRjobNutrients.py



```
from mrjob.job import MRJob

import re

import csv

WORD_RE = re.compile(r"[\w']+")
f = open("test.csv")
reader = csv.reader(f)


class MRWordFreqCount(MRJob):



    def mapper(self, _, line):
        temp = True
        for row in reader:
           if temp:
              temp = False
           else:
              yield(float(row[11]), row[0])

    def combiner(self, word, counts):

        yield (word, counts)



    def reducer(self, word, counts):

        s = sorted(word, reverse=True)[:10]
        for weights, names in zip(s, counts):
           yield (names, weights)



if __name__ == '__main__':

     MRWordFreqCount.run()
```

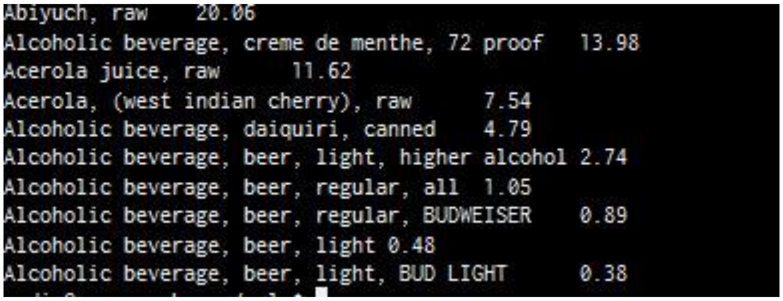
Results: The data set we used is like below:



```
name,weight,measure,protein-gm,protein-unit,protein-value,fat-gm,fat-unit,fat-value,carb-gm,carb-unit,carb-value
"Abiyuch, raw",114.0,0.5 cup,1.5,g,1.71,0.1,g,0.11,17.6,g,20.06
"Acerola juice, raw",242.0,1.0 cup,0.4,g,0.97,0.3,g,0.73,4.8,g,11.62
"Acerola, (west indian cherry), raw",98.0,1.0 cup,0.4,g,0.39,0.3,g,0.29,7.69,g,7.54
"Alcoholic beverage, beer, light",29.5,1.0 fl oz,0.24,g,0.07,0.0,g,0.00,1.64,g,0.48
"Alcoholic beverage, beer, light, BUD LIGHT",29.5,1.0 fl oz,0.25,g,0.07,0.0,g,0.00,1.3,g,0.38
"Alcoholic beverage, beer, light, BUDWEISER SELECT",29.5,1.0 fl oz,0.2,g,0.06,0.0,g,0.00,0.87,g,0.26
"Alcoholic beverage, beer, light, higher alcohol",356.0,12.0 fl oz,0.25,g,0.89,0.0,g,0.00,0.77,g,2.74
"Alcoholic beverage, beer, light, low carb",29.5,1.0 fl oz,0.17,g,0.05,0.0,g,0.00,0.73,g,0.22
"Alcoholic beverage, beer, regular, all",29.7,1.0 fl oz,0.46,g,0.14,0.0,g,0.00,3.55,g,1.05
"Alcoholic beverage, beer, regular, BUDWEISER",29.8,1.0 fl oz,0.36,g,0.11,0.0,g,0.00,2.97,g,0.89
"Alcoholic beverage, creme de menthe, 72 proof",33.6,1.0 fl oz,0.0,g,0.00,0.3,g,0.10,41.6,g,13.98
"Alcoholic beverage, daiquiri, canned",30.5,1.0 fl oz,0.0,g,0.00,--,g,--,15.7,g,4.79
"Alcoholic beverage, daiquiri, prepared-from-recipe",30.2,1.0 fl oz,0.06,g,0.02,--,g,--,6.94,g,2.10
"Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 100 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 80 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 86 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,0.1,g,0.03
"Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 90 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 94 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, gin, 90 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, rum, 80 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, vodka, 80 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,--,g,--
"Alcoholic beverage, distilled, whiskey, 86 proof",27.8,1.0 fl oz,0.0,g,0.00,--,g,--,0.1,g,0.03
"Alcoholic beverage, liqueur, coffee with cream, 34 proof",31.1,1.0 fl oz,2.8,g,0.87,15.7,g,4.88,20.9,g,6.50
"Alcoholic beverage, liqueur, coffee, 53 proof",34.8,1.0 fl oz,0.1,g,0.03,0.3,g,0.10,46.8,g,16.29
"Alcoholic beverage, liqueur, coffee, 63 proof",34.8,1.0 fl oz,0.1,g,0.03,0.3,g,0.10,32.2,g,11.21
"Alcoholic beverage, malt beer, hard lemonade",335.0,11.2 fl oz,0.0,g,0.00,--,g,--,10.07,g,33.73
"Alcoholic beverage, pina colada, canned",32.6,1.0 fl oz,0.6,g,0.20,7.6,g,2.48,27.6,g,9.00
"Alcoholic beverage, pina colada, prepared-from-recipe",31.4,1.0 fl oz,0.42,g,0.13,1.88,g,0.59,22.66,g,7.12
"Alcoholic beverage, rice (sake)",29.1,1.0 fl oz,0.5,g,0.15,0.0,g,0.00,5.0,g,1.46
"Alcoholic beverage, tequila sunrise, canned",31.1,1.0 fl oz,0.3,g,0.09,0.1,g,0.03,11.3,g,3.51
"Alcoholic beverage, whiskey sour",30.4,1.0 fl oz,0.0,g,0.00,0.03,g,0.01,13.17,g,4.00
"Alcoholic beverage, whiskey sour, canned",30.8,1.0 fl oz,0.0,g,0.00,--,g,--,13.4,g,4.13
```

Output:

The top 10 food which has height fat value.



```
Abiyuch, raw    20.06
Alcoholic beverage, creme de menthe, 72 proof   13.98
Acerola juice, raw      11.62
Acerola, (west indian cherry), raw      7.54
Alcoholic beverage, daiquiri, canned    4.79
Alcoholic beverage, beer, light, higher alcohol 2.74
Alcoholic beverage, beer, regular, all  1.05
Alcoholic beverage, beer, regular, BUDWEISER    0.89
Alcoholic beverage, beer, light 0.48
Alcoholic beverage, beer, light, BUD LIGHT      0.38
```

Discussion: Abiyuch has the highest fat value like 20.06, the surprising results are alcohol contents have highest fat value which is very dangerous.

Conclusion: I conclude that the highest cholesterol contents are a few juices and alcohol contents main food division that have high cholesterol is the alcohol. So for further assistance we can also analyze reducing this data in to similar groups and then conclude the list of food that has high or low fat value.