Python Project 06 Research Papers Analysis

Assignment Overview

- Files
- Lists and Tuples

Assignment Background

Life is full of wonders and uncertainties that motivates us to find answers. However, not everybody is qualified to answer such questions. We need to have a medium where various like-minded individuals with experience of certain topics to share and contribute their acquired knowledge via research articles. One well known medium to publish these articles is through research journals. These journals are centered on certain topics or categories (Biology, Computer Science, etc.), where some categories are more popular than others. This popularity is measured by their citation rate or impact factor.

For this project, the program reads files with the citation data of each category and the impact factor of various journals, print the average citation rate of the top 20 categories and plot the top 10 journal impact factor. The data used for this project was extracted from InCites Journal Citation Reports (https://jcr.incites.thomsonreuters.com/JCRLandingPageAction.action).

Project Specifications

1. You must implement the following functions:

   a) **open_file()** prompts the user for a filename to read data from. An error message should be shown if the file cannot be opened. This function will loop until it receives proper input and successfully opens the file. It returns a file pointer.

   b) **read_journal_file(fp)** reads the file object containing journal names and their impact factor. We are only interested in the journal name (str), total number of citations (int), and the impact factor (float). It returns a list of tuples, sorted by the impact factor (descending order). Only read the first 30 characters for the journal name. **Make sure to remove all commas from all strings with numeric values. See Notes and Hints. Make sure to check that the entries in citations and impact factor are of type int and float respectively. If it is not a valid entries don't include it in the list.**

   c) **read_category_file(fp)** reads the file object containing the citation data from over 200 categories. We are only interested in the category (str), number of journals (int), and total number of citations (int). Only read the first 30 characters for the category. For each

category, you need to calculate the average citation per journal (float). It returns a list of tuples containing the category, the number of journals, the total number of citations and the average citation per journal, sorted by category alphabetically (ascending order). **Make sure to remove all commas from all strings with numeric values. See Notes and Hints**

d) **display_table(data)** receives the list of tuples with the citation data per category. It prints the data and an extra row with the totals of each column separated from the table by a row of 85 `'-'`. It returns nothing. The formatting for the rows is: "{:30s}{:>10,d}{:>18,d}{:>25,.3f}". Each row has a total of 85 characters. The table headings should have the following formatting:

```
Title ("Citation Data of the Top 20 Categories"):
                          "{:^85s}"
Headings ('Category','Journals','Total Citations','Citation per
Journal'):
               "{:30s}{:>12s}{:>18s}{:>25s}"
```

e) **sort_data(data, column)** This function receives the list of tuples with the citation data per category and the column index to sort by. **The column index must start at 0**. Use the itemgetter function to sort by the indicated column. It returns the list sorted by the selected column. If sorted by category then the table should be sorted alphabetically in ascending order, otherwise it should be sorted in descending order.

f) **prepare_plot_data(data)** takes the top 10 journals with the highest impact factor and separates them into two lists (names and impact factor). The impact factor needs to be of type float. It returns a tuple of two lists.

g) **plot_data(name, data)** plots a bar chart of the top 10 journals with the highest impact factor. **This function is already provided for this project.**

h) **main()** This function will call all other functions in this project.

- Firstit will open and read both files (category and journal files).
- Then it prompts the user for a column index and will re-prompt for a column index until it is valid value has been entered (all digits and between 1 and 4).
- The list with the citation data per category needs to be sorted by the specified column and then display the sorted list. Only the first 20 categories will be displayed.
- Then, the user should be asked if they want to plot data: the pssible answers are yes or no. Capitalization does not matter.
- If yes, sorts and plots the journal impact factor data. First it extracts the name labels and impact factor of the journal citation data and then calls the `plot_data()` function provided.

Notes and Hints

1. The replace function is useful to remove unwanted characters from a string.
2. When reading the citation and journal files, use the reader function from csv module to read each line of the file because some values have commas.

    import csv

    fp = open(filename, ‘r’)

    csv.reader(fp)

3. The try-except structure will help determine whether the current column is numeric. You can try to convert a string to a float number. If Python raises a ValueError, it means that converting the string to a float type is not possible.

Sample Output:

read_journal_file function:

```
fp = open('journal_impact_small.csv','r')

instructor_data = [('REVIEWS OF MODERN PHYSICS', 41133, 33.177), ('NEW
ASTRONOMY REVIEWS', 922, 6.154), ('Nonlinear Analysis-Hybrid Syst',
812, 3.192), ('Microfluidics and Nanofluidics', 4089, 2.537),
('EXTREMOPHILES', 2718, 2.346), ('JOURNAL OF COMPUTING IN CIVIL ',
1541, 1.855), ('Journal of Neurosurgery-Pediat', 2567, 1.757), ('WORLD
BANK RESEARCH OBSERVER', 784, 1.667), ('Archaeological and
Anthropolog', 301, 1.636), ('SKELETAL RADIOLOGY', 4318, 1.527),
('ARCHAEOMETRY', 2112, 1.364), ('Molecular & Cellular Toxicolog', 299,
1.24), ('Applied Geophysics', 417, 0.804), ('Fixed Point Theory', 304,
0.581), ('China Communications', 275, 0.424)]
student_data = read_journal_file(fp)

assert instructor_data == student_data
```

read_category_file function test:

```
fp = open('category_impact_small.csv','r')
student_data = read_category_file(fp)
```

```
instructor_data = [('ACOUSTICS', 32, 138295, 4321.71875),
('AGRICULTURE, MULTIDISCIPLINARY', 57, 170336, 2988.3508771929824),
('BIOCHEMISTRY & MOLECULAR BIOLO', 289, 3273965, 11328.598615916955),
('CHEMISTRY, MULTIDISCIPLINARY', 163, 2825242, 17332.77300613497),
('ENGINEERING, ENVIRONMENTAL', 50, 510092, 10201.84), ('ENGINEERING,
GEOLOGICAL', 35, 76977, 2199.342857142857), ('ENVIRONMENTAL SCIENCES',
225, 1412031, 6275.693333333334), ('ETHNIC STUDIES', 15, 11308,
753.8666666666667), ('GEOGRAPHY, PHYSICAL', 49, 191491,
3907.9795918367345), ('GEOLOGY', 47, 102891, 2189.1702127659573),
('GERIATRICS & GERONTOLOGY', 49, 171259, 3495.081632653061), ('HEALTH
CARE SCIENCES & SERVICE', 88, 272255, 3093.806818181818), ('MATERIALS
SCIENCE, COATINGS & ', 18, 209367, 11631.5), ('MATERIALS SCIENCE,
TEXTILES', 23, 35426, 1540.2608695652175), ('PSYCHOLOGY, APPLIED', 79,
173846, 2200.5822784810125)]
assert instructor_data == student_data
```

sort_data function test:

```
student_data = [('ONCOLOGY', 213, 1634966, 7675.896713615023),
('CHEMISTRY, MEDICINAL', 59, 425363, 7209.542372881356),
('BIOTECHNOLOGY & APPLIED MICROBIOLOGY', 161, 1103236,
6852.39751552795), ('DEVELOPMENTAL BIOLOGY', 41, 273038,
6659.463414634146), ('BEHAVIORAL SCIENCES', 51, 305160,
5983.529411764706), ('MEDICINE, RESEARCH & EXPERIMENTAL', 124, 694043,
5597.120967741936), ('PHYSICS, MATHEMATICAL', 53, 283825,
5355.188679245283), ('METALLURGY & METALLURGICAL ENGINEERING', 73,
360924, 4944.164383561644), ('NUCLEAR SCIENCE & TECHNOLOGY', 32,
149291, 4665.34375), ('MARINE & FRESHWATER BIOLOGY', 104, 399530,
3841.6346153846152), ('AGRONOMY', 83, 237099, 2856.614457831325),
('VETERINARY SCIENCES', 138, 277519, 2011.0072463768115),
('SOCIOLOGY', 142, 178756, 1258.8450704225352), ('HISTORY & PHILOSOPHY
OF SCIENCE', 44, 22128, 502.90909090909093)]

instructor_data = [('AGRONOMY', 83, 237099, 2856.614457831325),
('BEHAVIORAL SCIENCES', 51, 305160, 5983.529411764706),
('BIOTECHNOLOGY & APPLIED MICROBIOLOGY', 161, 1103236,
6852.39751552795), ('CHEMISTRY, MEDICINAL', 59, 425363,
7209.542372881356), ('DEVELOPMENTAL BIOLOGY', 41, 273038,
6659.463414634146), ('HISTORY & PHILOSOPHY OF SCIENCE', 44, 22128,
502.90909090909093), ('MARINE & FRESHWATER BIOLOGY', 104, 399530,
3841.6346153846152), ('MEDICINE, RESEARCH & EXPERIMENTAL', 124,
694043, 5597.120967741936), ('METALLURGY & METALLURGICAL ENGINEERING',
73, 360924, 4944.164383561644), ('NUCLEAR SCIENCE & TECHNOLOGY', 32,
149291, 4665.34375), ('ONCOLOGY', 213, 1634966, 7675.896713615023),
('PHYSICS, MATHEMATICAL', 53, 283825, 5355.188679245283),
('SOCIOLOGY', 142, 178756, 1258.8450704225352), ('VETERINARY
SCIENCES', 138, 277519, 2011.0072463768115)]

assert instructor_data == sort_data(student_data,0)
```

prepare_plot_data function test:

```
student_data = [('Psychological Science in the Public Interest', 858,
19.286), ('Cell Metabolism', 21343, 17.303),('JOURNAL OF NUCLEAR
MEDICINE', 22728, 5.849),('CANCER', 62200, 5.649),('BRAIN PATHOLOGY',
4403, 5.256),('AMERICAN ECONOMIC REVIEW', 35805,
3.833),('INTERNATIONAL JOURNAL OF FOOD MICROBIOLOGY', 22247,
3.445),('DRUG SAFETY', 4104, 3.206),('PSYCHOLOGICAL ASSESSMENT', 7886,
2.901),('DRUGS & AGING', 2827, 2.61),('Journal of Diabetes
Investigation', 966, 2.294),('Journal of Real-Time Image Processing',
341, 1.564),('CROP SCIENCE', 15892, 1.55),('JOURNAL OF RADIATION
RESEARCH', 2071, 1.536),('Clinical Nursing Research', 563,
1.359),('Economics of Energy & Environmental Policy', 108,
1.172),('International Journal of Speech-Language Pathology', 485,
0.985),('AGROFORESTRY SYSTEMS', 2306, 0.91),('CREATIVITY RESEARCH
JOURNAL', 1712, 0.881),('Information Technology for Development', 258,
0.857),('INTERNATIONAL JOURNAL OF COMPUTATIONAL FLUID DYNAMICS', 506,
0.772),('NORWEGIAN JOURNAL OF GEOLOGY', 770, 0.69),('ASTRONOMY &
GEOPHYSICS', 134, 0.256)]

instructor_names = ['Psychological Science in the P', 'Cell
Metabolism', 'JOURNAL OF NUCLEAR MEDICINE', 'CANCER', 'BRAIN
PATHOLOGY', 'AMERICAN ECONOMIC REVIEW', 'INTERNATIONAL JOURNAL OF FOOD
', 'DRUG SAFETY', 'PSYCHOLOGICAL ASSESSMENT', 'DRUGS & AGING']

instructor_impact = [19.286, 17.303, 5.849, 5.649, 5.256, 3.833,
3.445, 3.206, 2.901, 2.61]

assert (instructor_names, instructor_impact) ==
prepare_plot_data(student_data)
```

Test Case 1:

```
Please enter a valid filename: category_impact_2017.csv

Please enter a valid filename: journal_impact_2017.csv

Column number to sort data (1-category, 2-journals, 3-citations, 4-average
citations): 1
```

Citation Data of the Top 20 Categories

Category	Journals	Total Citations	Citation per Journal
ACOUSTICS	31	174,802	5,638.774
AGRICULTURAL ECONOMICS & POLIC	17	24,021	1,413.000
AGRICULTURAL ENGINEERING	14	166,334	11,881.000
AGRICULTURE, DAIRY & ANIMAL SC	60	192,794	3,213.233
AGRICULTURE, MULTIDISCIPLINARY	57	210,711	3,696.684
AGRONOMY	87	287,102	3,300.023
ALLERGY	27	127,991	4,740.407
ANATOMY & MORPHOLOGY	21	65,760	3,131.429
ANDROLOGY	6	8,410	1,401.667

```
ANESTHESIOLOGY                          31          201,325              6,494.355
ANTHROPOLOGY                            85          126,983              1,493.918
AREA STUDIES                            68           38,910                572.206
ASTRONOMY & ASTROPHYSICS                66        1,071,345             16,232.500
AUDIOLOGY & SPEECH-LANGUAGE PA          25          100,231              4,009.240
AUTOMATION & CONTROL SYSTEMS            61          350,086              5,739.115
BEHAVIORAL SCIENCES                     51          356,259              6,985.471
BIOCHEMICAL RESEARCH METHODS            79          797,638             10,096.684
BIOCHEMISTRY & MOLECULAR BIOLO         293        3,625,819             12,374.809
BIODIVERSITY CONSERVATION               57          207,782              3,645.298
BIOLOGY                                 85          491,775              5,785.588
-------------------------------------------------------------------------------
TOTAL                                1,221        8,626,078            111,845.400

Do you want to plot the journal data (yes/no)? no
```

Test Case 2:

```
Please enter a valid filename: xxxx

Error with the file. Please enter a valid filename: category_impact_2016.csv

Please enter a valid filename: journal_impact_2017

Error with the file. Please enter a valid filename: journal_impact_2016.csv

Column number to sort data (1-category, 2-journals, 3-citations, 4-average
citations): 4

                     Citation Data of the Top 20 Categories
Category                           Journals   Total Citations     Citation per Journal
MULTIDISCIPLINARY SCIENCES                64        2,803,793                43,809.266
PHYSICS, CONDENSED MATTER                 67        1,344,772                20,071.224
CHEMISTRY, PHYSICAL                      146        2,865,201                19,624.664
PHYSICS, ATOMIC, MOLECULAR & C            36          705,786                19,605.167
CHEMISTRY, MULTIDISCIPLINARY             166        3,088,211                18,603.681
ELECTROCHEMISTRY                          29          531,031                18,311.414
PHYSICS, PARTICLES & FIELDS               29          474,927                16,376.793
ASTRONOMY & ASTROPHYSICS                  63          979,492                15,547.492
NANOSCIENCE & NANOTECHNOLOGY              87        1,332,720                15,318.621
CHEMISTRY, ORGANIC                        59          774,862                13,133.254
PHYSICS, APPLIED                         148        1,875,869                12,674.791
ENGINEERING, ENVIRONMENTAL                49          619,251                12,637.776
BIOCHEMISTRY & MOLECULAR BIOLO           290        3,435,913                11,847.976
MATERIALS SCIENCE, COATINGS &             19          224,479                11,814.684
PHYSICS, MULTIDISCIPLINARY                79          915,427                11,587.684
ENERGY & FUELS                            92        1,031,892                11,216.217
PHYSICS, NUCLEAR                          20          218,773                10,938.650
AGRICULTURAL ENGINEERING                  14          150,733                10,766.643
CELL BIOLOGY                             190        2,044,775                10,761.974
MATERIALS SCIENCE, MULTIDISCIP           275        2,957,270                10,753.709
-------------------------------------------------------------------------------------
TOTAL                                  1,922       28,375,177               315,401.678
```

```
Do you want to plot the journal data (yes/no)? n
Incorrect answer! Enter yes/no

Do you want to plot the journal data (yes/no)? no
```

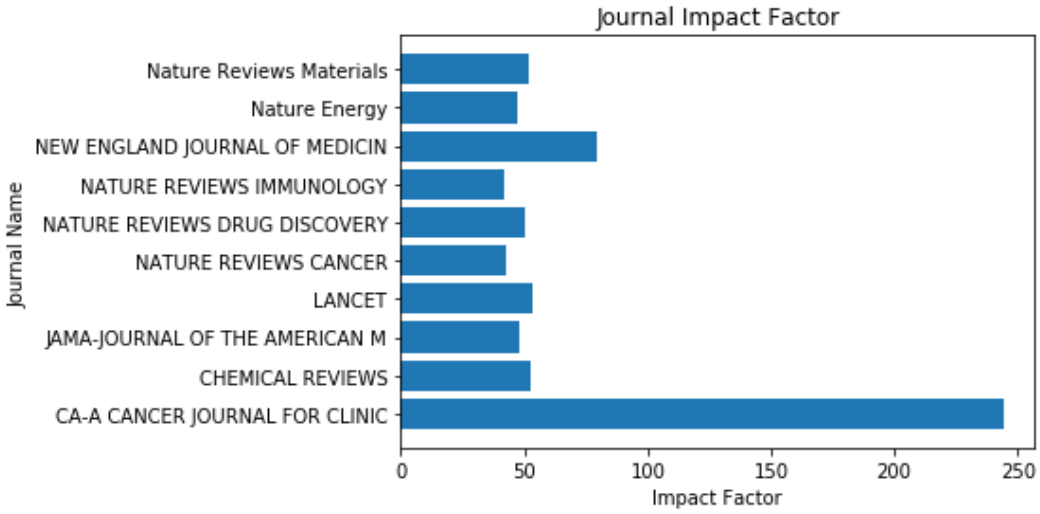
Test Case 3: Plot Test

```
Please enter a valid filename: category_impact_2017.csv

Please enter a valid filename: journal_impact_2017.csv

Column number to sort data (1-category, 2-journals, 3-citations, 4-average
citations): 3
```

Citation Data of the Top 20 Categories

Category	Journals	Total Citations	Citation per Journal
BIOCHEMISTRY & MOLECULAR BIOLO	293	3,625,819	12,374.809
CHEMISTRY, MULTIDISCIPLINARY	171	3,468,236	20,282.082
MATERIALS SCIENCE, MULTIDISCIP	285	3,451,318	12,109.888
CHEMISTRY, PHYSICAL	147	3,187,930	21,686.599
MULTIDISCIPLINARY SCIENCES	64	3,132,708	48,948.562
NEUROSCIENCES	261	2,346,383	8,989.973
CELL BIOLOGY	190	2,134,575	11,234.605
PHYSICS, APPLIED	146	2,094,241	14,344.116
ONCOLOGY	223	1,931,396	8,660.969
ENVIRONMENTAL SCIENCES	242	1,893,304	7,823.570
ENGINEERING, ELECTRICAL & ELEC	260	1,636,339	6,293.612
NANOSCIENCE & NANOTECHNOLOGY	92	1,579,362	17,166.978
PHARMACOLOGY & PHARMACY	261	1,571,415	6,020.747
PHYSICS, CONDENSED MATTER	67	1,490,204	22,241.851
MEDICINE, GENERAL & INTERNAL	155	1,456,323	9,395.632
BIOTECHNOLOGY & APPLIED MICROB	161	1,323,552	8,220.820
CLINICAL NEUROLOGY	197	1,303,928	6,618.924
IMMUNOLOGY	155	1,280,207	8,259.400
ENERGY & FUELS	97	1,278,572	13,181.155
SURGERY	200	1,206,541	6,032.705
TOTAL	3,667	41,392,353	269,886.997

```
Do you want to plot the journal data (yes/no)? yes
```


Journal Impact Factor
Nature Reviews Materials
Nature Energy
NEW ENGLAND JOURNAL OF MEDICIN
NATURE REVIEWS IMMUNOLOGY
NATURE REVIEWS DRUG DISCOVERY
NATURE REVIEWS CANCER
LANCET
JAMA-JOURNAL OF THE AMERICAN M
CHEMICAL REVIEWS
CA-A CANCER JOURNAL FOR CLINIC
Journal Name
0
50
100
150
200
250
Impact Factor