CS433/CS603 Programming Assignment Data Cleaning

Please follow the instructions to complete the following Python programs. For each program, please also provide your own testing cases. Please complete them in either Jupter notebook or with .py file and submit your programs and running results.

Read the article at https://realpython.com/python-data-cleaning-numpy-pandas/#tidying-up-fields-in-the-data and do further cleaning work on the following two .csv files:

BL-Flicker-Images-Book.csv file

For the Title column, only keep the title part, remove all of the other parts. If the title is inside a bracket, remove the bracket which encloses the title. If there are multiple …, only keep the words before the first one. If there are multiple periods (.), only keep words before the first one. The following shows the sample results after the cleaning for title column:

Walter Forbes. [A novel.] By A. A -> *Walter Forbes.*

Love the Avenger. By the author of â€œAll for Greed.â€ [The dedication signed: A. A. A., i.e. Marie Pauline Rose, Baroness Blaze de Bury.] -> *Love the Avenger.*

[The World in which I live, and my place in it. By E. S. A. [i.e. Letitia Willgoss Stone.] Edited by ... J. H. Broome.] -> *The World in which I live, and my place in it.*

A Satyr against Vertue. (A poem: supposed to be spoken by a Town-Hector. [By John Oldham. The preface signed: T. A.]) -> *A Satyr against Vertue.*

An Account of the many and great Loans, Benefactions and Charities, belonging to the City of Coventry ... A new edition. [The dedication signed: AB, CD, EF, GH, &c. By Edward Jackson and Samuel Carte.] -> *An Account of the many and great Loans, Benefactions and Charities, belonging to the City of Coventry ...*

For the Author and Contributors columns, only keep the first author, remove all of the other parts. If the author or the contributor has other auxiliary information, remove them. title is inside a bracket, remove the bracket which encloses the title. If there are multiple …, only keep the words before the first one. If there are multiple periods (.), only keep words before the first one. For all of the names, they should only have the first letter of the first, middle and last names be capital letter, all of the remaining letters should be small case. The following shows the sample results after the cleaning for title column:

Author

A., J.|A., J. -> *A., J.*

AAR, Ermanno - pseud. [i.e. Luigi Giuseppe Oronzo Mariano Raffaele Francesco Fortunato Felice de Simone.] -> *Aar, Ermanno*

Contributors:

CARTE, Samuel.|JACKSON, Edward - Rector of Southam, and CARTE (Samuel) -> *Carte, Samuel.*

After cleaning the above columns, save the cleaned data back into a new .csv file.

olympics.csv file

Drop all of the countries which don't have gold medals, and save the cleaned one back into a new .csv file.

Output the top five gold medal countries.

Output the countries that have all three kinds of medals for both summer and winter games.

You may use previous unit learned concepts on multi-level indexing, pivoting, stacking to rearrange the original .csv file to convert summer and winter as secondary level index within the country index, and then generate a new .csv file which will have country as the primary index, then summer and winter as secondary index, and columns of gold, silver and bronze medals.