Shell commands

In this task, you are required to explore and wrangle the data in the file “***covid19-cable-broadcast-labeled.csv***”, which contains 44,643 transcript paragraphs that were collected from various programs in cable and broadcast news networks, e.g., World News Tonight on ABC and Anderson Cooper 360 Degrees on CNN. These transcript paragraphs have been manually annotated according to their relevance to COVID-19. The file contains different variables to describe each collected transcript paragraph, as described below.



|  |  |
| --- | --- |
| **Column Name**  | **Description**  |
| *network*  | Cable and broadcast news networks such as ABC, CNN, FOX, etc.  |
| *program*  | Programs aired on cable and broadcast news networks, such as “worldnewstonight” (the program World News Tonight on ABC)  |
| *date*  | The date when the corresponding program aired  |
| *speech\_turn*  | Each transcript characterises the flow of coverage for a particular program at a particular date. This generally requires speakers taking speech turns. This column encodes this value.  |
| *paragraph\_sequence*  | Sometimes a given turn takes multiple paragraphs. This column gives the paragraph sequence number in each turn  |
| *paragraph*  | The textual content of the paragraph  |
| *category*  | The categorical label provided by humans to indicate a paragraph’s relevance to COVID-19, i.e., *covid\_direct*, *covid\_indirect*, and *non\_covid*  |

Please note that you are only allowed to use shell commands as you would run in Linux shell, Mac terminal, or Cygwin, to tackle this task. Using other utilities or tools such as PowerShell is NOT allowed.

Task

1. What is the *date* range of the collected paragraphs? Please note that the file is not guaranteed to be sorted and Nulls (NA and empty values) should not be considered.
2. How many times was “Australia” mentioned in the column *paragraph*? When was the first mention of the term “*Australia*” in the column *paragraph*? Please note that the first mention of a term refers to the paragraph which not only contains the term but also the corresponding *date* value of the paragraph is earliest in the dataset, and the term to be searched is case sensitive.
3. How many unique values are there in the *program* column? Write commands to list the top 5 most frequent *program* values in the dataset (i.e., the top 5 programs with the largest number of paragraphs in the dataset)? Please also write commands to list the 5 least frequent *program* values in the dataset.
4. Write commands to count the number of times the words “*COVID-19*” and “*pandemic*” appeared in the column *paragraph* of the dataset (ignoring cases). If a term appears twice in a text value, it should be counted twice. How many paragraphs contain each of the two words and how many paragraphs include both of these two terms in the dataset? (also ignoring cases)
5. In the following, let’s focus on certain columns contained in the file. Write commands to only keep columns that are in the following list:

*○ network
○ program ○ date
○ paragraph ○ category*

In particular, please only keep the paragraphs satisfying the following: (i) whose date is after March 31, 2020; (ii) whose *network* value is “cbs”; and (ii) whose *category* value is “covid\_direct”. How many paragraphs are left? Export the selected data to a new file named “***filtered\_paragraphs.csv***”. Please ensure that the file “filtered\_paragraphs.csv” contains data from the selected columns as well as the column names