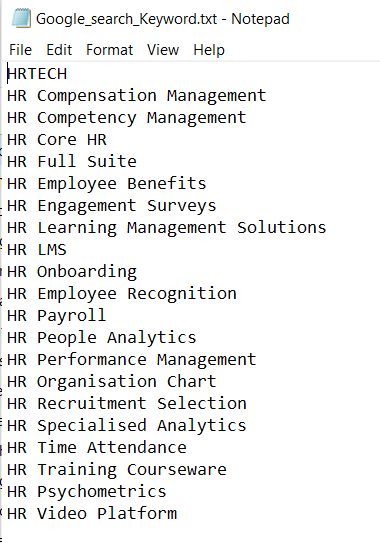
**Main Program to run:**

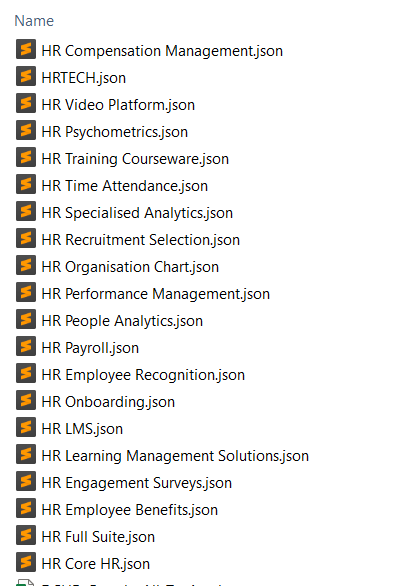
1. Run batch file - **GoogleRSS\_Scrape.bat**
   1. This is to download all googlerss news related to the keywords selected in the “input\Google\_search\_Keyword.txt”.
   2. Need to input any keywords to this text file for desirable keywords to scrape
2. Run batch file - **Create\_Google\_news\_dataset.bat**
   1. This is to create dataset used by the Microsoft Power BI tools for Data Visualisation purpose
   2. Optional to input to the stopwords into the “input\stopwords.txt” file for undesirable words to be display in the wordcloud diagram
3. Run the “**DCHR-Dashboard-GoogleNews.pbix**” to open out this Microsoft Power BI dashboard file.
   1. This will display the wordcloud and related data from the excel dataset created by step 2.

**For GoogleRSS HR news scrapping procedures:**

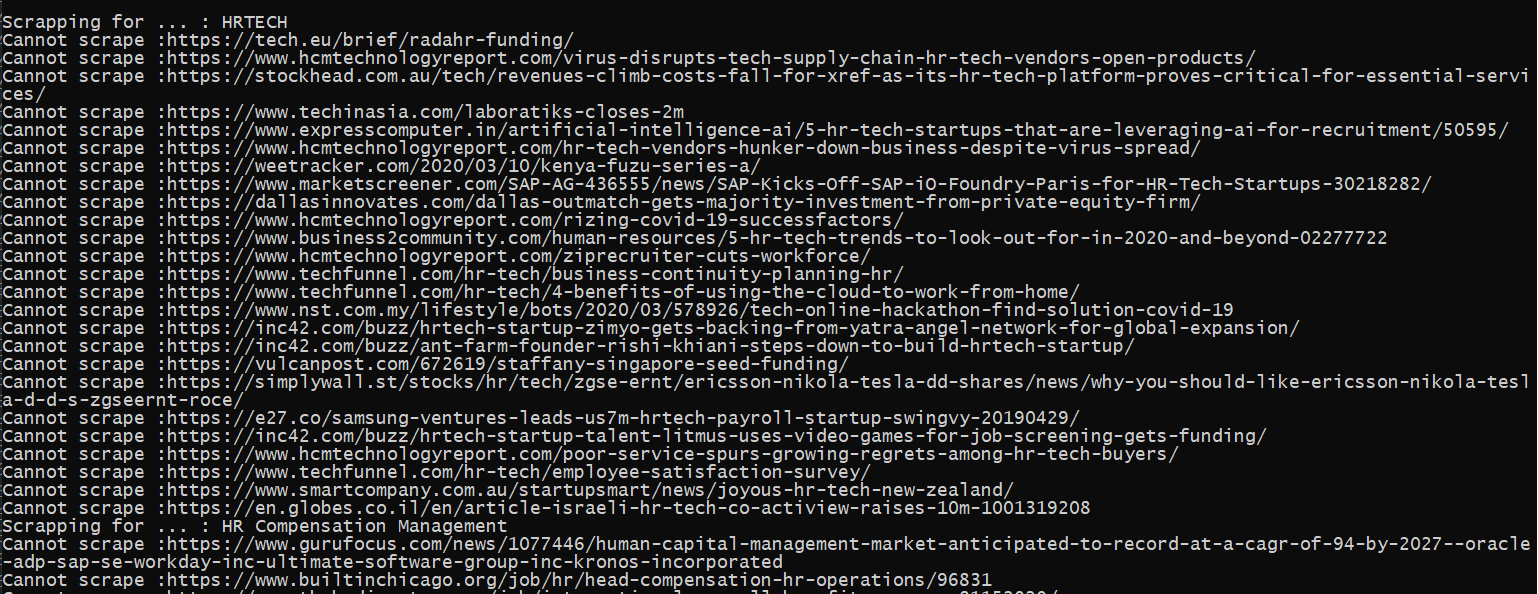
1. Run batch file - GoogleRSS\_Scrape.bat
   1. This scrapper will scrape news from Google RSS site
   2. It will read the keywords to scrape from input folder “input\Google\_search\_Keyword.txt”
   3. For new keywords to scrap, need to input into this text file line by line



1. After each keyword is scrapped, the data is save into json format with the keyword names. Eg scrapping keyword HRTECH, the data will be saved into data\HRTECH.json.

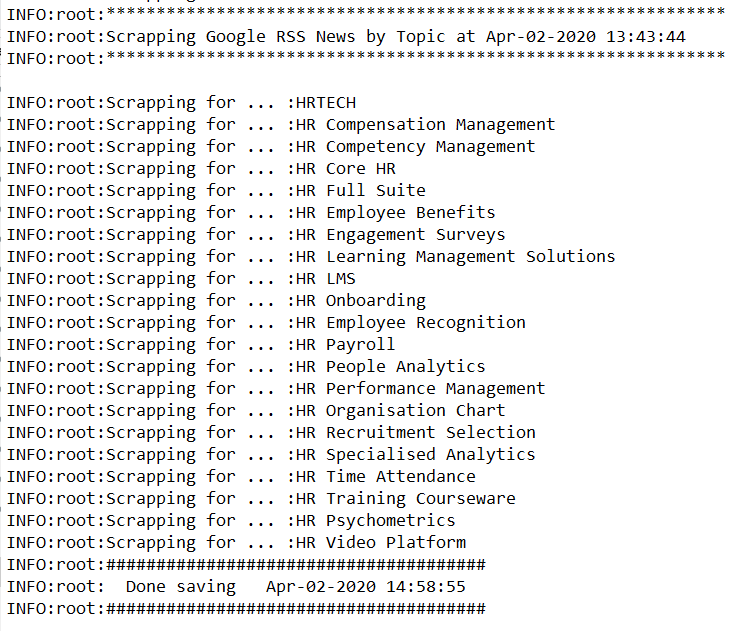


1. Output console:



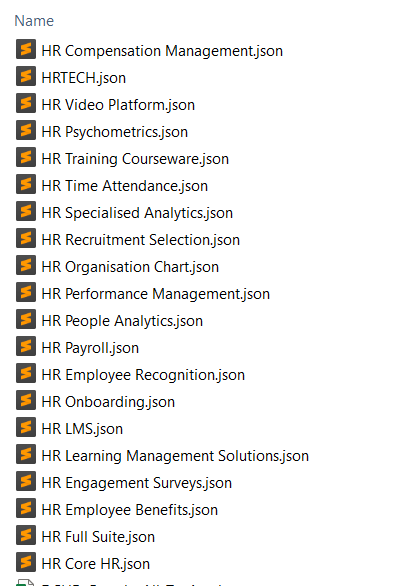
The “Cannot scrape …” message means it will not scrape the site where it has restriction of scrapping, hence it will skip the site and carry on to the next available sites.

1. Log files will be created in the log folder “log\DCHR\_Scrape\_Google\_news\_by\_topic\_info.log”:

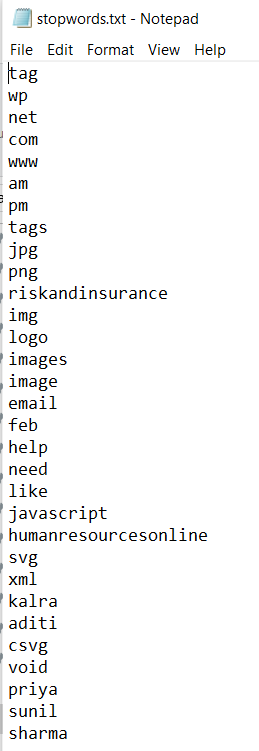


**For Creating GoogleRSS HR news dataset procedures:**

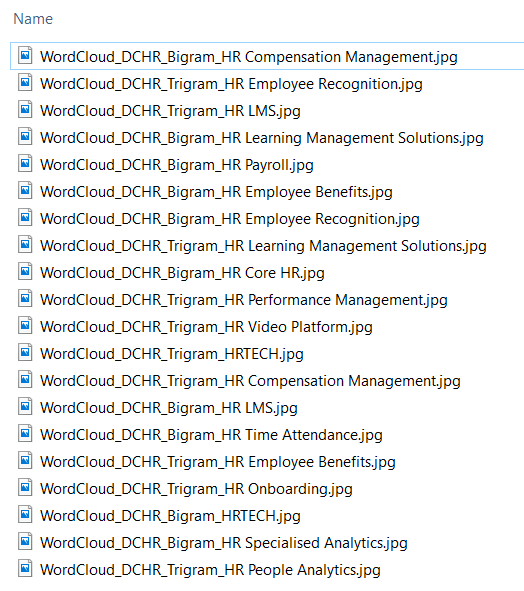
1. Run batch file - Create\_Google\_news\_dataset.bat
   1. This will create dataset in excel format from all the json files created from the GoogleRSS\_Scrape.bat run in data folder



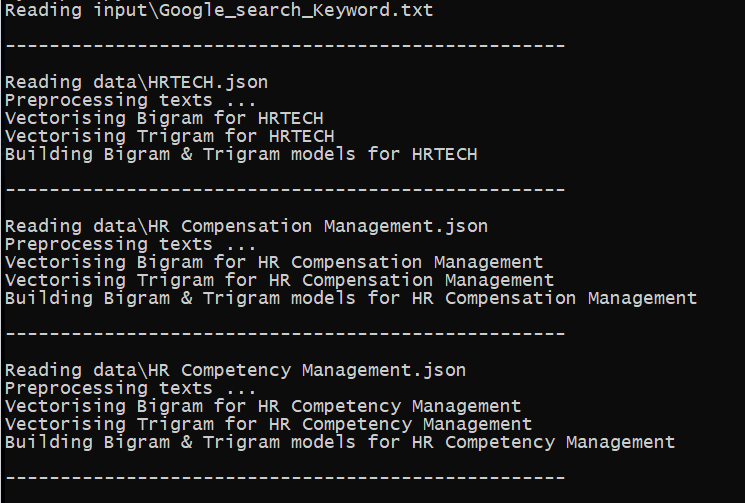
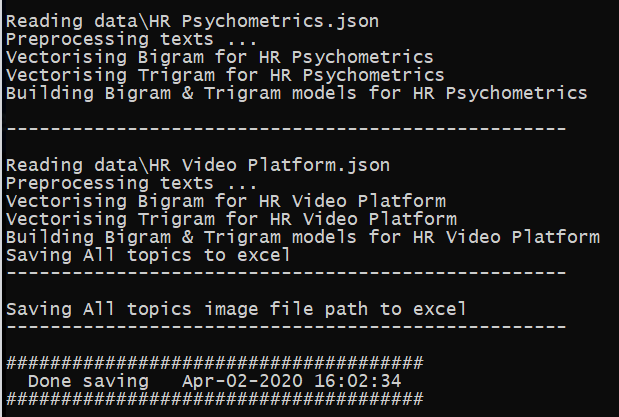
* 1. It will read the stopwords text file from to scrape from input folder “input\stopwords.txt”. This contains the words that are undesirable to show up in the wordcloud.
  2. For new stopwords, need to input into this text file line by line



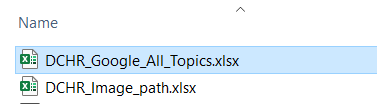
1. For every HR category data, bigram (2 words wordcloud) and trigram (3 words wordcloud) diagram will be generated in jpg format store in the diagram folder. Eg diagram\ WordCloud\_DCHR\_Bigram\_HRTECH.jpg.



1. All these jpg will not be used as input to the Power BI due to slow processing for displaying and difficult for selection reference. These jpg files will be converted to base64 format and save into excel file as another database for Power BI to point to. These data will be saved into folder data/DCHR\_Image\_path.xlsx.
2. Output console will also shows the progress of the program execution:

1. Each of the HR catergory json files data will be processed and converted to a excel file saved into data/DCHR\_Google\_All\_Topics.xlsx.

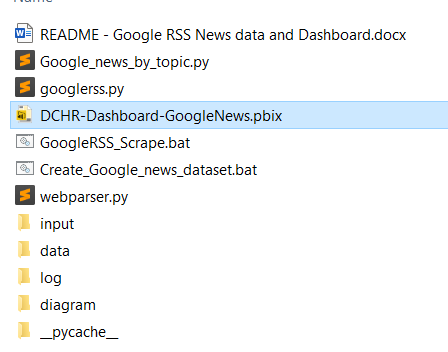


1. Log files will be created in the log folder “log\DCHR\_Google\_new\_by\_topic\_info.log”:

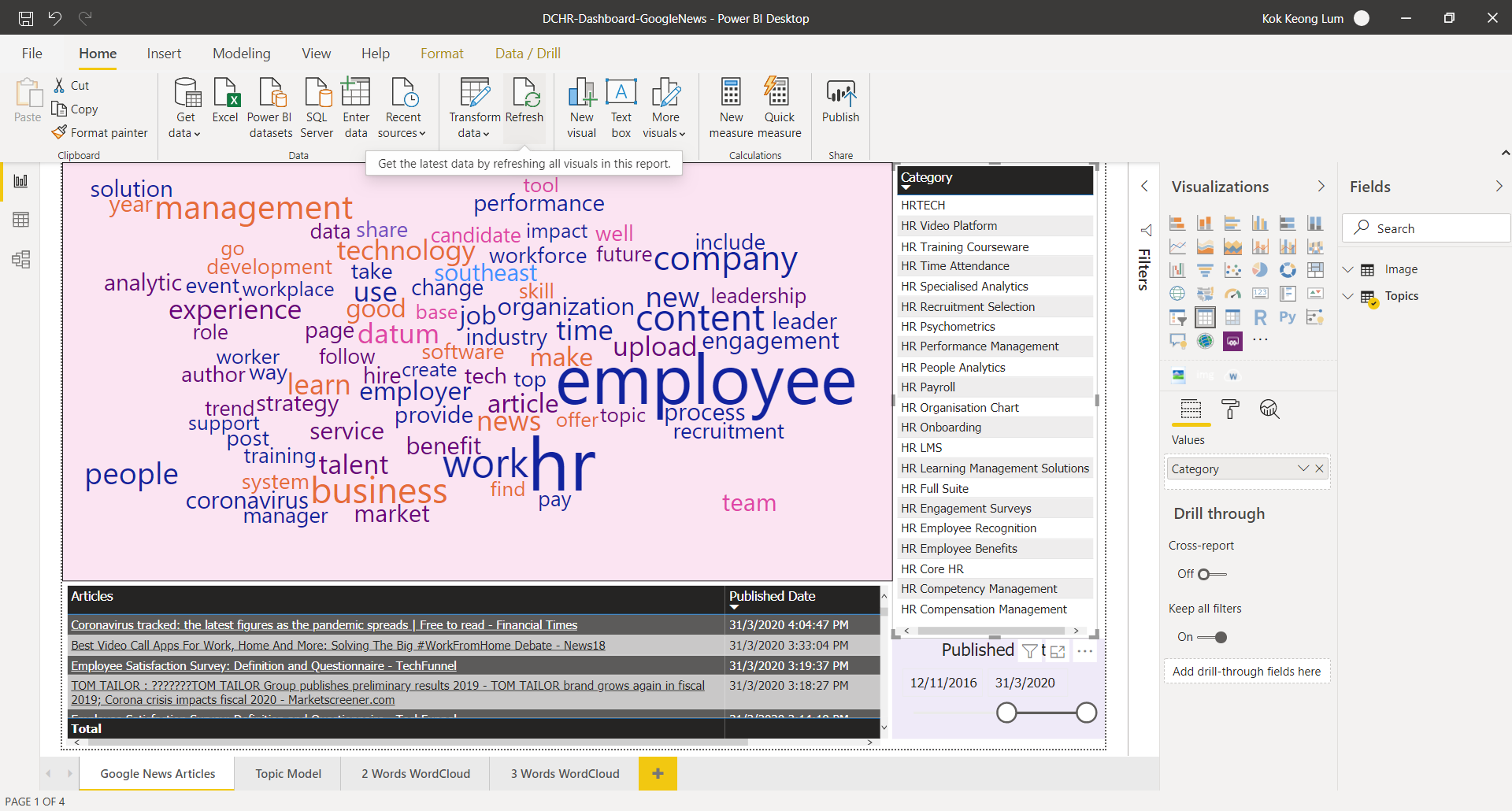


**Using Microsoft Power BI Dashboard:**

1. Run or double click the “DCHR-Dashboard-GoogleNews.pbix” file to open up the Microsoft Power BI application



1. First Page on the Power BI Dashboard “Google News Article” shows the overall dashboard of the Google News for all the HR Category. Click on the “Refresh” button on the Menu for updating the data from the latest updates.



1. Top Right Window - “Category” Selection
   1. In this window, it allows one to selected the desire category. Once selected, the action will trigger the rest of the two windows, “WordCloud” and “Articles & Published Date” to updates information with reference to the selected category
   2. When one category is selected, the rest will be greyed out and the “WordCloud” and “Articles & Published Date” windows will refresh and updated
   3. To return to orginal state, click on the selected category word to deselection



1. Top Left Window – “WordCloud” Display
2. Bottom Left Window “Published Date” Selection
3. Bottom Left Window – “Articles & Published Date” Selections