Quotation Tool notebook – help pages

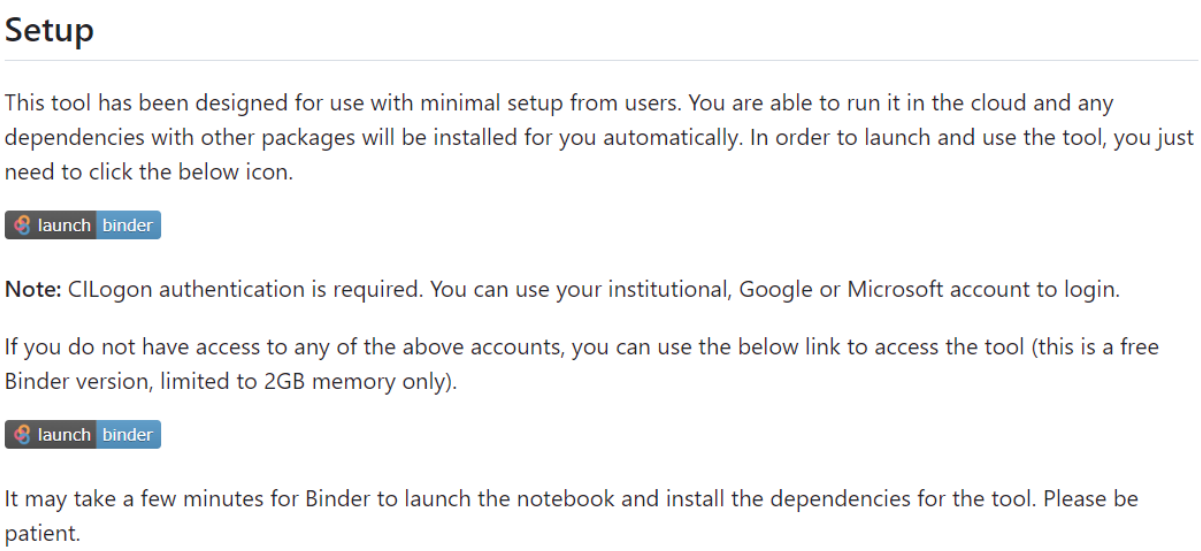
Introduction

The ATAP Quotation Tool is a Jupyter notebook containing code that was adapted and developed (with permission) from the GenderGapTracker by the Sydney Informatics Hub (SIH) in collaboration with the Sydney Corpus Lab as part of the Australian Text Analytics Platform (ATAP) project. The tool is designed to identify and extract quotes from newspaper texts. In addition to extracting the quotes, the tool also provides information about who the speakers are, identifies the reporting expressions used (e.g. *say*, *tell*), and classifies entities (e.g. as person, organisation, location) both regarding the speakers and the quoted content. Users can download the results as a spreadsheet for further analysis.

(Note: if you are unfamiliar with how to use Jupyter Notebooks, have a look at this guide.)

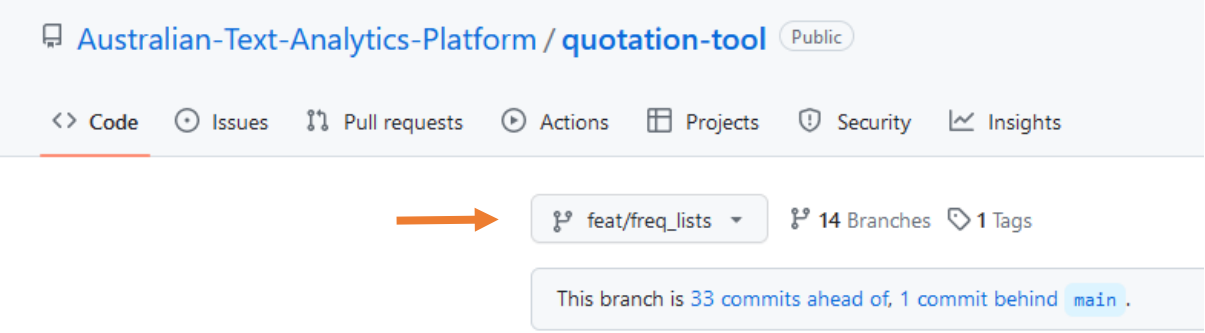
Getting started

The tool is available on GitHub where you can launch the tool on Jupyter Notebook via Binder by clicking on one of the 'launch binder' buttons:



The access to the ATAP Binderhub (i.e., the first 'launch binder' button) requires CILogon authentication, which supports the single sign-on (SSO) method with most (Australian or international) institutional login credentials or a Google/Microsoft account. If you have trouble authenticating, please refer to the CILogon troubleshooting guide. If you have access to software that supports Jupyter Notebooks, you can also clone the Github repository and use the notebook locally (i.e., without Internet connection) on your own computer.

In late 2023 we created a minor update of this notebook with a new feature integrated in the downloaded spreadsheet. This version is currently available by changing the "main" branch to the "feat/freq_lists" branch before launching the notebook via Binder.



Alternatively, you can access this notebook version via this URL: https://github.com/Australian-Text-Analytics-Platform/quotation-tool/tree/feat/freq_lists

This version of the notebook will generate the full results (similar to the earlier version) as well as providing frequency lists for most of the categories (e.g., verbs, speakers; see notebook itself for a full list of frequency list categories). These frequency lists are available as separate sheets within the Excel spreadsheet.

Overview of Tool

If you have already read the blog post introducing this tool or are familiar with the tool, you can skip this general overview section and will find the tool explanation from here onwards.

The quotation tool is designed to identify and extract quotes (cited or projected) from newspaper texts. Since the tool uses a combination of syntactic and heuristic rules to extract quotes, it is able to identify quotes whether they are marked by quoting or projecting verbs (e.g. *said*) or not.

The tool can also use Named-Entity Recognition to identify and classify the sources of these quotes and the entities within these quotes (as people, organisations, etc.). This is useful for answering a number of questions such as those related to representation of voices/sources (e.g. Who is cited the most/least? Who is not cited at all?) and those related to quoted content (e.g. What kind of information is sourced from others?).

The tool can also identify verbs often used to cite quotations (e.g. *say*, *tell*, *add*, *claim*). This is particularly useful to those interested in the variation of reporting verbs.

Once the tool has processed the files, it will display the first few identified and extracted quotes (and entities) in a table. An example of this preview table is shown in Table 1 below, which displays the identified quote along with information such as the speaker and their entity type, the entity name and type of the entities identified within the quote and the quoting/projecting verb (if there is one). The type of quote is described on the basis of the various components of the quote construction (Q = quotation mark, S = speaker, V = verb, C = content) and their linear order.

Table 1. Preview of first few identified and extracted quotes

text_id	text_name	quote_id	quote	quote_index	quote_entities	speaker	speaker_index	speaker_entities	verb	verb_index	quote_token_count	quote_type	is_floating_quote
:b2ac4cab7493e7	1-Morrison_defends_ministerial_appointments	0	In hindsight these arrangements were unnecessary	(487, 535)	[]	he	(538, 540)	[]	said	(541, 545)	6	QCQSV	False
:b2ac4cab7493e7	1-Morrison_defends_ministerial_appointments	1	The risk of ministers becoming incapacitated, ...	(755, 900)	[]	he	(903, 905)	[]	wrote	(906, 911)	26	QCQSV	False
:b2ac4cab7493e7	1-Morrison_defends_ministerial_appointments	2	The Australian people have been let down, they...	(1055, 1120)	[(Australian, NORP)]	she	(1123, 1126)	[]	said	(1127, 1131)	12	QCQSV	False
:b2ac4cab7493e7	1-Morrison_defends_ministerial_appointments	3	people were on ventilators and people were dyi...	(1479, 1648)	[(Morrison, PERSON)]	he	(1651, 1653)	[]	told	(1654, 1658)	34	CSV	False
:b2ac4cab7493e7	1-Morrison_defends_ministerial_appointments	4	he was also unaware that his portfolio was tak...	(1765, 1897)	[(Morrison, PERSON)]	Former finance minister Simon Birmingham	(1719, 1759)	[(Simon Birmingham, PERSON)]	said	(1760, 1764)	26	SVC	False

The tool allows you to save and download the complete table of results as an Excel worksheet (.xslx format) for further analysis.

It is also possible to preview all the identified quotes and entities in individual files. In this preview (see Figure 1 below), identified quotes and entities are presented in bold face and labelled accordingly (e.g. as quote/speaker or as a specific entity type such as PERSON, NORP, GPE – see legend to Figure 2 for abbreviations). You can download the visualisation such as the one shown in Figure 1 as an html file.

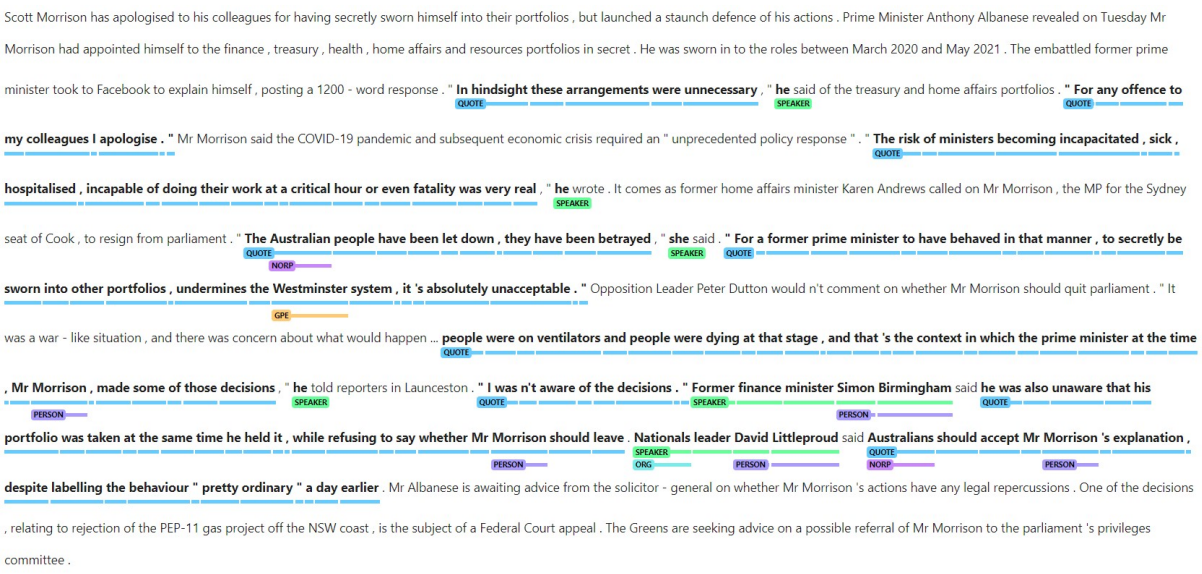


Figure 1. Screenshot of preview showing identified entities and quotes

The tool also allows you to visualise the top named entities identified in the quotes and/or the top entity types among identified speakers as bar graphs – an example is shown in Figure 2 below.

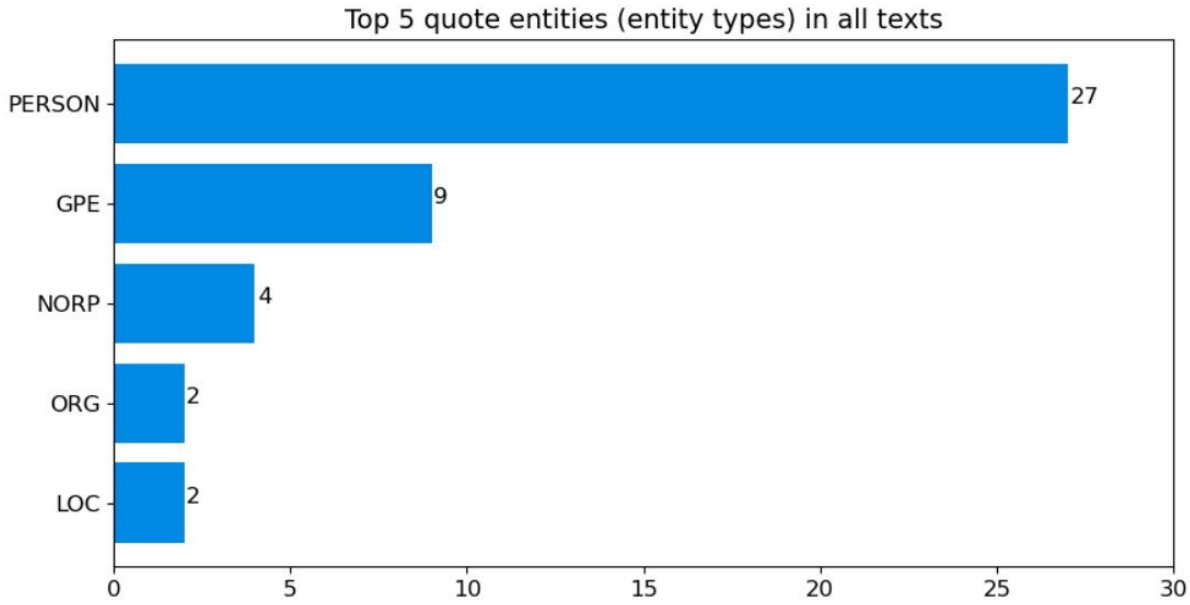


Figure 2. Bar graph showing the top five entities identified within quotes across the whole corpus/dataset.

(PERSON = People, including fictional, GPE = Countries, cities, states, NORP = Nationalities or religious or political groups, ORG = Companies, agencies, institutions, etc., LOC = Non-GPE locations, mountain ranges, bodies of water)

You can choose to visualise the top entities for the whole corpus/dataset or individual files within the corpus/dataset. Other options for the visualisation include whether to display the entity names and/or types, and the number of top entities to display (i.e. in multiples of five). You can also choose to save the graphs (based on the set parameters) as jpg files.

Setup

Before you begin, you need to import the QuotationTool package and the necessary libraries and initiate them to run in this notebook.

1. Execute the cell:

```
[ ]: # import the QuotationTool
     from extract_display_quotes import QuotationTool, DownloadFileLink

     # initialize the QuotationTool
     qt = QuotationTool()
```

2. Once completed, you should get a message saying, “Finished loading”:

```
[1]: # import the QuotationTool
     from extract_display_quotes import QuotationTool, DownloadFileLink

     # initialize the QuotationTool
     qt = QuotationTool()

     [nltk_data] Downloading package punkt to /home/jovyan/nltk_data...
     [nltk_data]   Unzipping tokenizers/punkt.zip.
     Loading spaCy language model...
     This may take a while...
     Finished loading.
```

Load the data

This notebook will allow you to extract quotes directly from a text file (or a number of text files). Alternatively, you can also extract quotes from a "text" column inside your excel spreadsheet (for the current versions, please ensure the uploaded spreadsheet contains exact column headers as shown):

	text_name	text
2	text1	Facebook and Instagram, which Facebook owns, followed up in the evening, announcing that Trump wouldn't be able to post for 24 hours following two violations of its policies. The White House did not immediately offer a response to the actions. While some cheered the platforms' response, experts noted that these actions follow years of hemming and hawing regarding Trump and his supporters spreading dangerous misinformation and encouraging violence that contributed to Wednesday's events.
3	text2	(CBC News) Republican lawmakers and previous administration officials had begged Trump to give a statement to his supporters to quell the violence. He posted his video as authorities struggled to take control of a chaotic situation at the Capitol that led to the evacuation of lawmakers and the death of at least one person. Lawmakers, world leaders condemn chaos at the U.S. Capitol while some call for Trump's removal Trudeau says Canadians 'deeply disturbed' by violence in Washington D.C.

1. Execute the cell:

```
[ ]: # upload the text files and/or excel spreadsheets onto the system
     display(qt.upload_box)
     print('Uploading large files may take a while. Please be patient.')
     print('\033[1mPlease wait and do not press any buttons until the progress bar appears...\033[0m')
```

2. Click 'Upload your files'.

```
[2]: # upload the text files and/or excel spreadsheets onto the system
     display(qt.upload_box)
     print('Uploading large files may take a while. Please be patient.')
     print('\033[1mPlease wait and do not press any buttons until the progress bar appears...\033[0m')
```

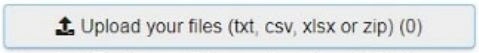
Upload your files (txt, csv, xlsx or zip) (0)

```
Uploading large files may take a while. Please be patient.
Please wait and do not press any buttons until the progress bar appears...
```

A window should appear prompting you to select txt files, or a single csv file, xlsx file, or zip folder.

3. Click 'Open' after you've selected the file(s) you want to upload.
4. The tool should start loading the selected file(s). Please be patient and do not press any buttons until the progress bar appears; uploading large files may take a while. Once completed, you get a message saying, "Finished uploading files" and another message stating the number of files that were uploaded (e.g., "100 text documents are loaded for tagging"):

```
[2]: # upload the text files and/or excel spreadsheets onto the system
     display(qt.upload_box)
     print('Uploading large files may take a while. Please be patient.')
     print('\033[1mPlease wait and do not press any buttons until the progress bar appears...\033[0m')
```




```
The total size of the upload is 0.38 MB.
Reading uploaded files...
This may take a while...
100%|██████████| 100/100 [00:00<00:00, 93435.15it/s]
Finished uploading files.
100 text documents are loaded for tagging.
```

5. Execute the following cell to view a snippet of the content of the files you've uploaded:

```
[ ]: # specify the number of rows you wish to display
     n=5

     # display a preview of the pandas dataframe
     qt.text_df.head(n)
```

By default, you will preview the first 5 rows of the extracted quotes in table format:

```
[3]: # specify the number of rows you wish to display
     n=5

     # display a preview of the pandas dataframe
     qt.text_df.head(n)
```

[3]:

	text_name	text	text_id
0	2021_01_17_LaurenLancaster	Activists hold swim-in for trans inclusion at ...	dd181c3672dc3359353eda6497d585c3
1	2021_01_21_JulietteMarchant	NTEU appeals Tim Anderson Federal Court ruling...	cc8fe4275242db85d93392e5a705d4e0
2	2021_01_25_MaxShanahan	Health Minister rejects Invasion Day Covid pla...	29c99bc924bd2a1c4f744ff89f9d605b
3	2021_02_05_ShaniaOBrien	NUS condemns "horrific" assault on internation...	01aceb45444212877ad3c6b8a340ac85
4	2021_02_14_OliverPether	17 years on, activists continue to demand just...	26a7b1ff63231b9e8f181abf86827d40

You can adjust this by changing the number value in "n=5" in the code cell (to e.g., n=2, n=10, n=20) and re-execute the cell.

Extract the quotes

Once your texts have been uploaded, you can extract the quotes from the texts. You can also extract named entities from your text.

1. You can change the variables/parameters (for the types of entities you want the tool to extract and the preview of the results) in the cell shown below:

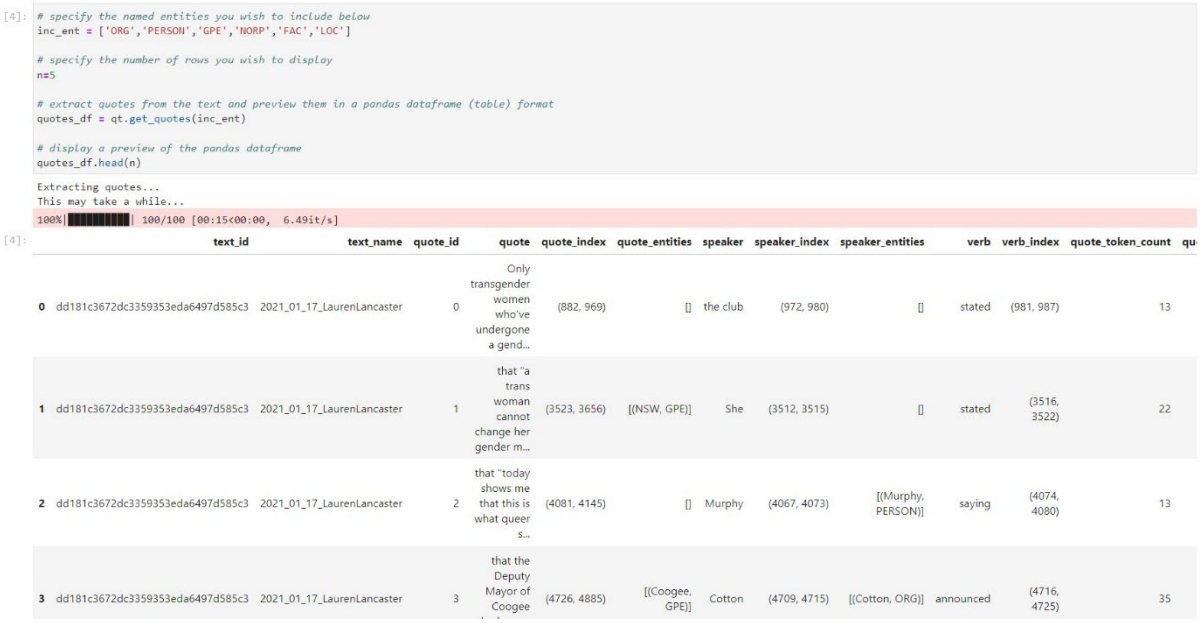
```
[ ]: # specify the named entities you wish to include below
     inc_ent = ['ORG','PERSON','GPE','NORP','FAC','LOC']

     # specify the number of rows you wish to display
     n=5

     # extract quotes from the text and preview them in a pandas dataframe (table) format
     quotes_df = qt.get_quotes(inc_ent)

     # display a preview of the pandas dataframe
     quotes_df.head(n)
```

2. Specify the entities you wish to include by modifying the values within the square brackets for “inc_ent”. By default, the notebook will attempt to identify entities (i.e., nouns or noun phrases) and label them as ORG, PERSON, GPE, NORP, FAC, or LOC (see notebook for an explanation of what these refer to). You can choose to delete any categories within the square brackets and only focus on identifying the ones of interest to you. Alternatively, you can add further entities (e.g. MONEY, PRODUCT, etc).
   Note: The available named entity types in spaCy may vary among different versions of pre-trained language models. In order to obtain a list of all NER types and their meanings, the updated notebook version (available at https://github.com/Australian-Text-Analytics-Platform/quotation-tool/tree/feat/freq_lists) contains a code cell that produces a list of all available entities of the loaded language model and their explanations prior to the code cell where you specify these entities.
3. Specify the number of rows you wish to display. By default, the notebook will display five rows (i.e., “n=5”). You can adjust this by changing the number/value for “n”.
4. Once you’re happy with the parameters, execute the cell to begin extracting the quotes and named entities. Please note that this can take some time, depending on the size of the dataset that is analysed.
5. Once completed, you should get a table similar to the one shown below:



```
[4]: # specify the named entities you wish to include below
inc_ent = ['ORG','PERSON','GPE','NORP','FAC','LOC']

# specify the number of rows you wish to display
n=5

# extract quotes from the text and preview them in a pandas dataframe (table) format
quotes_df = qt.get_quotes(inc_ent)

# display a preview of the pandas dataframe
quotes_df.head(n)

Extracting quotes...
This may take a while...
100%|██████████| 100/100 [00:15<00:00,  6.49it/s]
```

[4]:

	text_id	text_name	quote_id	quote	quote_index	quote_entities	speaker	speaker_index	speaker_entities	verb	verb_index	quote_token_count	qu
0	dd181c3672dc3359353eda6497d585c3	2021_01_17_LaurenLancaster	0	Only transgender women who've undergone a gend...	(882, 969)	[]	the club	(972, 980)	[]	stated	(981, 987)	13	
1	dd181c3672dc3359353eda6497d585c3	2021_01_17_LaurenLancaster	1	that "a trans woman cannot change her gender m...	(3523, 3656)	[(NSW, GPE)]	She	(3512, 3515)	[]	stated	(3516, 3522)	22	
2	dd181c3672dc3359353eda6497d585c3	2021_01_17_LaurenLancaster	2	that "today shows me that this is what queer s...	(4081, 4145)	[]	Murphy	(4067, 4073)	[(Murphy, PERSON)]	saying	(4074, 4080)	13	
3	dd181c3672dc3359353eda6497d585c3	2021_01_17_LaurenLancaster	3	that the Deputy Mayor of Coogee	(4726, 4885)	[(Coogee, GPE)]	Cotton	(4709, 4715)	[(Cotton, ORG)]	announced	(4716, 4725)	35	

   The table should contain quotes extracted by the notebook along with other information such as text id, text name (i.e., file name), speaker, speaker entity, verb, and so on (please refer to the notebook itself for an explanation of the information contained in each column of the table).

Display the quotes

Once you have extracted the quotes, you can see a preview of the quotes using spaCy's visualisation tool, displaCy.

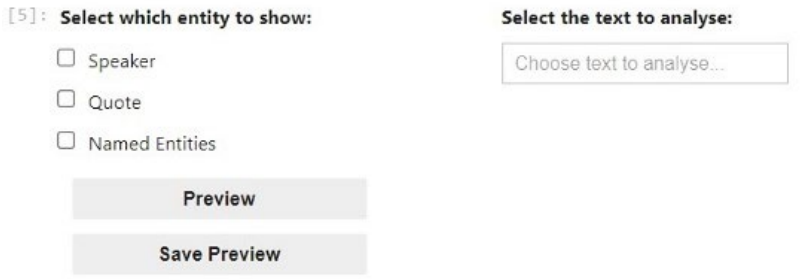
Displaying extracted quotes, speakers, and entities

1. Execute the cell:

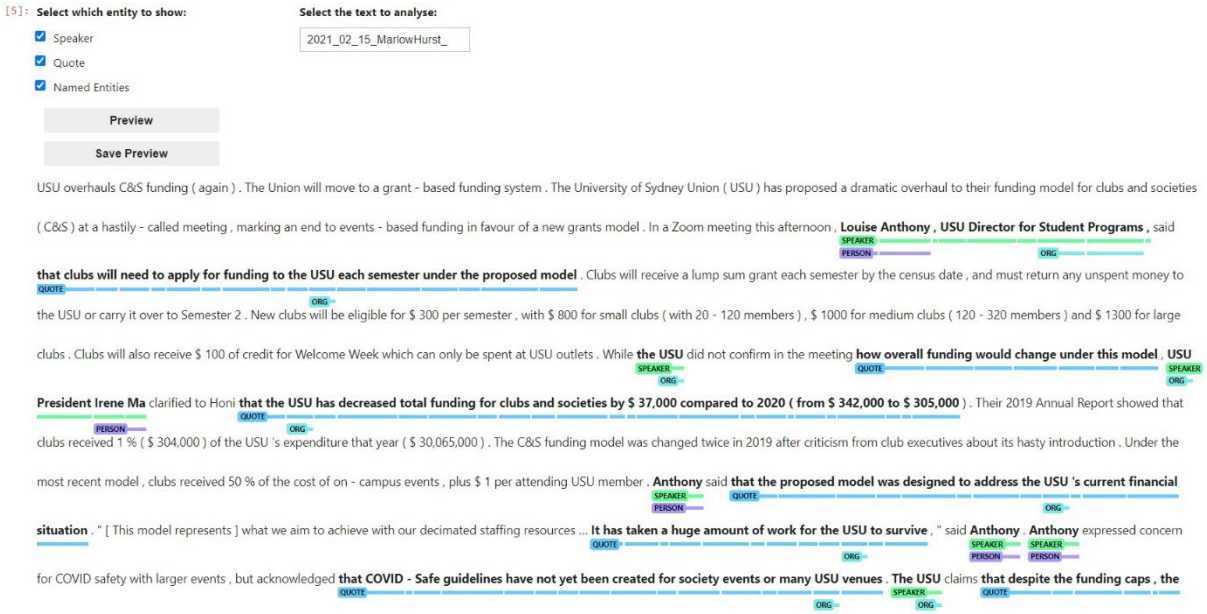
```
[ ]: # display a preview of the extracted quotes, speakers and entities within the text
     qt.analyse_quotes(inc_ent)
```

2. Once completed, you should get several widgets to adjust the settings for the preview of the extracted quotes, speakers, and entities:

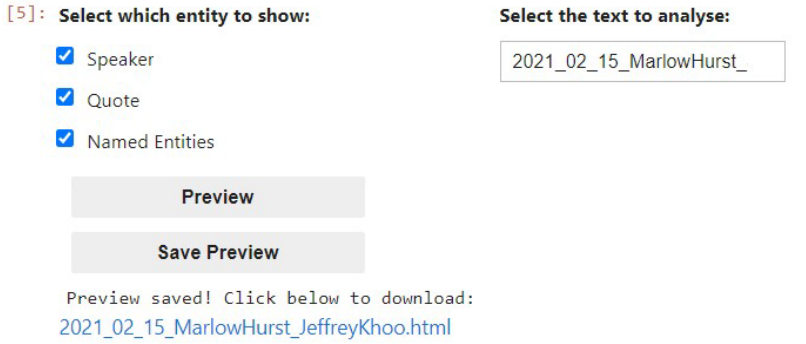
```
[5]: # display a preview of the extracted quotes, speakers and entities within the text
     qt.analyse_quotes(inc_ent)
```



3. Select one or more entities to show: Speaker, Quote, and Named Entities. *Speaker* is the identified source of the quote. *Quote* is the content identified as attributed to the source. *Named Entities* refers to the semantic categories (e.g., ORG, PEOPLE) that the tool assigns to the nouns or noun phrases that either represent the speaker or are mentioned within the identified quotes. (Entities that do not occur as a speaker or within a quote are not classified.)
4. Select which text within the corpus/dataset you want to look at. You can only select one text for preview.
5. Once you've adjusted the settings, click "Preview". The preview is useful for seeing how the tool works and the types of errors it may make. The text you've selected will be visualised along with the analysed entities labelled (i.e., Speaker, Quote, and/or Named Entities categories). For example:



6. Click “Save Preview” to save the result of this analysis. Click on the resulting link to save the visualisation as an html file to your computer:



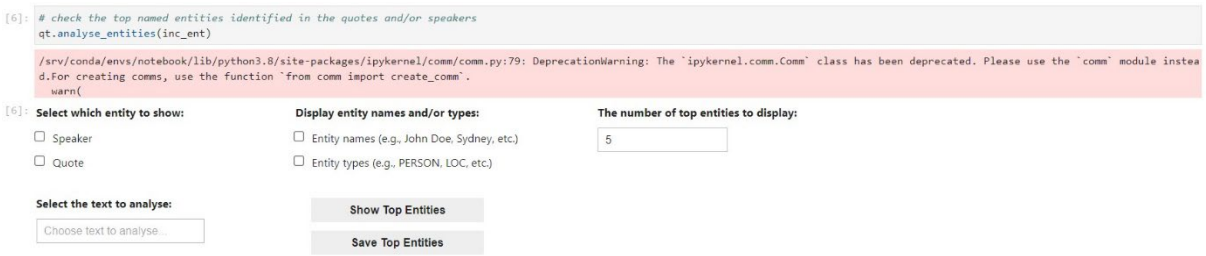
   By default, the html file should have the same filename (but different file extension) to the file you selected for analysis.
7. Repeat steps 2 and 5 for each text you want to examine (as well as step 6 for the ones you want to save). To choose a different text, first clear the currently selected text (e.g. by selecting/highlighting it and pressing “Delete” or “Backspace”). This enables you to select a different text for the preview. (Tip: typing part of the targeting text name will filter the dropdown list for a quicker selection.)

Displaying top named entities identified

The tool allows you to visualise the top named entities identified in the quotes and/or the top entity types among identified speakers as bar graphs.

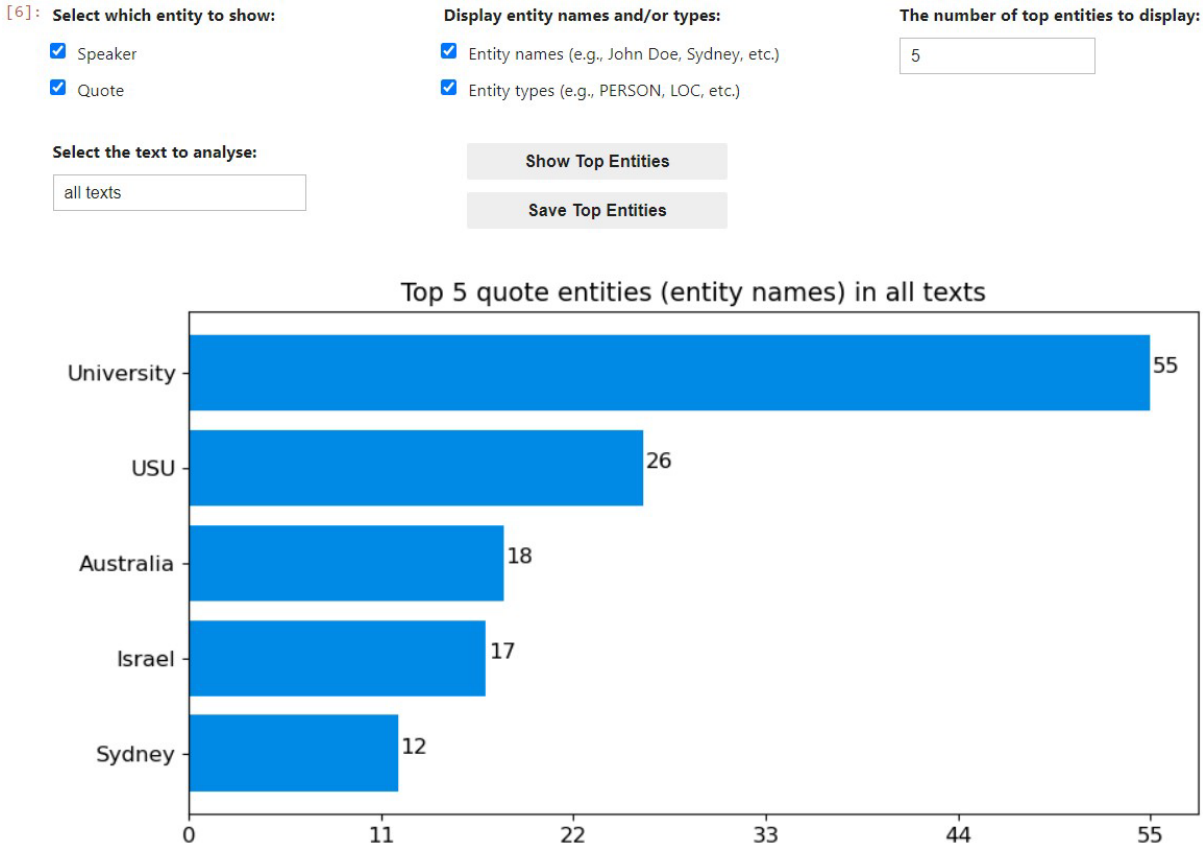
1. Execute the cell:

2. Once completed, you should get several widgets to adjust the settings for displaying the top named entities identified:



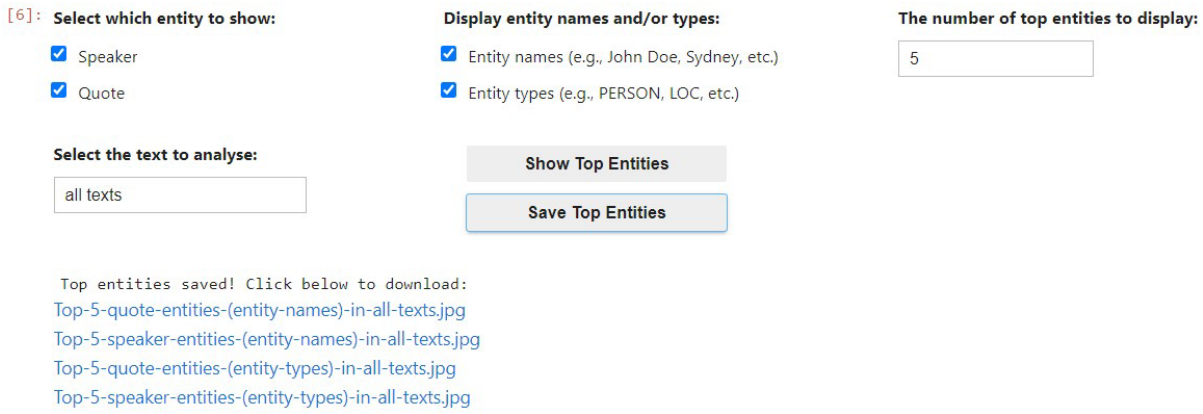
3. Select which type of entities (i.e., Speaker and Quote) you want to produce a bar graph for. You may select one or both of these options.
4. Select whether to produce bar graphs of entity names, entity types, or both.
5. Specify the number of top entities to display. By default, the notebook will only include the top five entities (i.e., “n=5”). You can adjust this by changing the number/value for “n”.

6. Select which text you want to analyse. You have the option to select individual texts within your corpus/dataset. You can also select the "all texts" options to examine the top entities across your entire corpus/dataset.
7. Once you're happy with the settings, click "Show Top Entities".
8. The tool will produce 1-4 bar graphs. For example:



If you select all options, the tool produces four separate graphs, showing the **names** of entities (within Quote and Speaker respectively) first, followed by the **types** of entities (within Quote and Speaker respectively). The 'top n quote entities (entity names)' graph shows the names of the entities that occur in quoted content (e.g. *Australia*); the 'top n speaker entities (entity names)' graph shows the names of the entities that are identified as source of the quote; the 'top n quote entities (entity types)' graph shows the types of the entities that occur in quoted content (e.g. ORG); the 'top n speaker entities (entity types)' graph shows the most frequent entity types within the source of the quote (e.g. PERSON).

9. Click "Save Top Entities" to save the result of this analysis.
10. Click on the resulting link(s) to the jpg file(s) to save it/them to your computer:

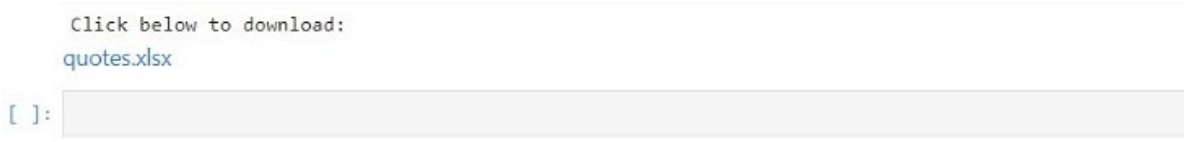


There should be a separate link for each of the bar graph produced.

11. If you're interested in examining specific texts in your corpus/dataset, repeat these steps for every text you want to analyse. (As before, you need to clear the current text selection first to enable you to select other texts, and type to filter if you know the name of target texts.)

Save the quotes

You can save the extracted quotes and all other information into an Excel spreadsheet and download it onto your computer by executing the code cell in this section. Note that the default file name is "quotes.xslx", which you can change in the code cell if you wish. Once you have done so, click on the resulting link to the xslx file to save it to your computer:



The Excel spreadsheet contains all of the quotes identified as well as information such as text id, text name, speaker, and so on (i.e., the same information shown in the preview produced during the Extract the quotes step).

Note: There may be an empty code cell at the end of the notebook, which you can ignore.

As a reminder, the minor update of the notebook (available in the feat/freq_lists branch) will generate the full results as well as frequency lists for most of the categories (e.g., verbs, speakers) within the downloaded spreadsheet as separate sheets). The steps to download this spreadsheet are the same.

Citing/Referencing Notebook

Citation: Jufri, S. & Sun, C. (2022). Quotation Tool (version1.0) [Jupyter notebook]. Australian Text Analytics Platform. https://github.com/Australian-Text-Analytics-Platform/quotation-tool

You can adjust the year, version number and URL in the above citation depending on the version of the notebook that you have used in your study.

If you are using this notebook in your research, please include the following statement or an appropriate variation thereof:

This study has utilised a notebook/notebooks developed for the Australian Text Analytics Platform (https://www.atap.edu.au) available at https://github.com/Australian-Text-Analytics-Platform/quotation-tool.

In addition, please inform ATAP (info@atap.edu.au) of publications and grant applications deriving from the use of any ATAP notebooks in order to support continued funding and development of the platform.

Acknowledgments

This Jupyter notebook and relevant python scripts were developed by the Sydney Informatics Hub (SIH) in collaboration with the Sydney Corpus Lab under the Australian Text Analytics Platform program and the HASS Research Data Commons and Indigenous Research Capability Program. These projects received investment from the Australian Research Data Commons (ARDC), which is funded by the National Collaborative Research Infrastructure Strategy (NCRIS).

Known Issues

The notebook has not been tested with very big data sets.