**User guide**

A screenshot of a cell phone

Description automatically generated

The only file to be edited is the ‘config.yml’ file in the config folder. It can be opened in any text editor like notepad++, atom etc. It can be opened in notepad like a normal text file, but the use of a better text editor is recommended to get a better sense of the indentation.

The program can be executed for multiple columns in multiple data files at one go. The only constraint is that all the data files should be located in the same folder/path. The following are the main parts of the file which need to be edited according to the requirements:

1. **path:** (line 2 in the figure above) The path needs to be changed to the location of the folder containing the data files. All the data files should be put only in this folder.
2. **name:** (line 9 in the figure above) This is the name of the ‘csv’ file in which data is present.
3. **id:** (line 10 in the figure above) This is the id column in the above csv data file, it has to be a unique identifier of rows in the dataset. If a unique identifier is not present, please create it manually before running the code.
4. **topic\_columns:** (lines 12-13 in the figure above) Enter the column names from the above csv data file (one in each line) on which analysis needs to be performed. The results from the ‘topic categorization’ will be generated for these columns. This is suitable mainly for the descriptive questions (long responses in multiple sentences) where the students write about various aspects of their life at UBC.
5. **senti\_columns:** (lines 15-16 in the figure above) Enter the column names from the above csv data file (one in each line) on which analysis needs to be performed. The results from the ‘sentiment analysis’ will be generated for these columns. This is suitable mainly for the descriptive questions (long responses in multiple sentences) where the students write about various aspects of their life at UBC, and their opinions might vary while writing about various aspects. If a column is entered here, both topic categorization and sentiment analysis will be performed as we are interested in how sentiments vary across various aspects.
6. **kp-unigram\_columns:** (line 18 in the figure above) Enter the column names from the above csv data file (one in each line) on which analysis needs to be performed. The results from the ‘key-word analysis’ will be generated for these columns. This is suitable mainly for the fact-based questions where the responses are short (few words or one sentence) but can be run on any question. Results generated will show few important words in the responses.
7. **kp-bigram\_columns:** (line 20 in the figure above) Enter the column names from the above csv data file (one in each line) on which analysis needs to be performed. The results from the ‘key-phrase analysis’ will be generated for these columns. This is suitable mainly for the fact-based questions where the responses are short (few words or one sentence) but can be run on any question. Results generated will show few important two-word phrases in the responses.

The above steps can be repeated for any number of files. As shown in the lines 22-32 in the above figure, the files can be added one after the other.

**Note:** If some kind of analysis is not required for any question, the name of analysis can be commented (by putting ‘#’ in front of the line) as shown in the lines 31-32 in the figure above, or it can be removed completely. Commenting it out is recommended as it makes it easier to edit if the analysis is required later.

**IMPORTANT: Add some data for training:** Please use historical and the current data to train the algorithm in order to develop the word clusters. To do this, simply add responses to open-ended questions in similar surveys in the `response` column in the `topic\full\_data\_train\_word2vec.csv` file. More data is better. See screenshot below for sample. It is recommended to add cleaned text (lower case, without stop words, etc). Functions in `cleaning\cleaning.py` can be referred to for this purpose.



**Stop Words:** If any word(s) are not desired in the output or if they don't add any value to the output, they can be added to the 'config/pair\_stopwords.txt' file.

**Flag Words:** If any words need to be added in the flagged words list, they can be added to the 'flag/flag\_words.txt' file.

**Note on the output:** PDF files are generated with bar graphs for each type of analysis. New folder called output will be generated and PDF files can be accessed in that folder.