Data Wrangling Exercise

Files

process_data.py
helikes.xlsx
identify_arguments.py
“She likes” vs “He likes”.pptx
shelikes.xlsx

Assumptions

- the dataset has trigrams followed by their frequency (separated by a space)
- the second word / token in the trigram is the verb / what the script should filter by (everything else is filtered out)
- the first trigram is on the fifth line of the file
- all of the trigrams with the specified verb are in the beginning and not mixed in
- the original file is readable
- the output file is writeable

External Libraries

The nltk package in Python is used to get the stop words (from nltk.corpus). No other external libraries have been used.

Filtering the Data

Run the process_data.py file using the following command –

```
python3 process_data.py -i [path/to/file] -o [path/to/new/file]
-v [verb]
```

The `-i`, `-o`, and `-v` are required.

`-i` is for the input file that has all of the data to be wrangled.
`-o` is for the output file to write to. If the output file has contents, it will be overwritten.
`-v` is for the verb in the trigrams that should be kept and should be passed in as a string.

The process_data.py script will go through the following steps to process the data –

1) Use the identify_arguments.py file to distinguish between the input file, the output file, and the verb.
2) Read in the data from the input file and add each line with the specified verb as the second word / token to a list of trigrams.
3) Iterate through the list to remove trigrams that have stop words as the third word / token since they don’t carry any meaning.
4) Write to the specified output file in the relative path (if specified) or in the current directory (if the relative path is unspecified). Each line in the output file will have a trigram and its corresponding frequency, separated by a space.

Other Output

Two numbers are printed in the console by the time process_data.py has finished running.

The first number printed is the number of trigrams (with corresponding frequencies) in the original file.

The second number printed is the number of trigrams (with corresponding frequencies) after removing stop words.

Excel Spreadsheets

The two Excel spreadsheets (helikes.xlsx and shelikes.xlsx) contain the manual filtering and sorting of the trigrams by frequency. Each sheet is more filtered than the previous sheet.

Visualizations

In “She likes” vs “He likes”.pptx, there are three visualizations. The first two visualizations are word clouds for “She likes” and “he likes” individually. The third visualization combines some of the most frequent words following “he likes” and “She likes”, and words that make the top of both lists are in the middle.